Khaled Elleithy
Tarek Sobh
Magued Iskander
Vikram Kapila
Mohammad A. Karim
Ausif Mahmood *Editors*

Technological Developments in Networking, Education and Automation

Springer

Technological Developments in Networking, Education and Automation

Khaled Elleithy · Tarek Sobh · Magued Iskander ·
Vikram Kapila · Mohammad A. Karim ·
Ausif Mahmood
Editors

Technological Developments in Networking, Education and Automation

Springer

Editors

Prof. Khaled Elleithy
University of Bridgeport
School of Engineering
University Avenue 221
06604 Bridgeport Connecticut
USA
elleithy@bridgeport.edu

Dr. Tarek Sobh
University of Bridgeport
School of Engineering
University Avenue 221
06604 Bridgeport Connecticut
USA
sobh@bridgeport.edu

Magued Iskander
Polytechnic University
Dept. Civil & Environmental
Engineering
Jay Street 333
11201 Brooklyn New York
USA
Iskander@poly.edu

Dr. Vikram Kapila
Polytechnic University
Dept. Civil & Environmental
Engineering
Jay Street 333
11201 Brooklyn New York
USA
vkapila@poly.edu

Dr. Mohammad A. Karim
Old Dominion University
Monarch Way 4111
23508 Norfolk Virginia
USA
mkarim@odu.edu

Ausif Mahmood
University of Bridgeport
School of Engineering
University Avenue 221
06604 Bridgeport Connecticut
USA
mahmood@bridgeport.edu

ISBN 978-90-481-9150-5 e-ISBN 978-90-481-9151-2
DOI 10.1007/978-90-481-9151-2
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010927951

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To our families
Their continuing support makes all our endeavors possible
And
To the memory of Professor Natalia Romalis
Her spirit will always be with us

Preface

This book includes the proceedings of the 2009 International Conferences on Telecommunications and Networking (TENE 09), Engineering Education, Instructional Technology, Assessment, and E-learning (EIAE 09), and Industrial Electronics, Technology & Automation (IETA 09). TENE 09, IETA 09 and EIAE 09 are part of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 09). The proceedings are a set of rigorously reviewed world-class manuscripts presenting the state of international practice in Industrial Electronics, Automation, Telecommunications, Networking, E-Learning, Instruction Technology, Assessment and Engineering Education.

CISEE 09 was a high-caliber research conference that was conducted online. CISSE 09 received 432 paper submissions and the final program included 220 accepted papers from more than 80 countries, representing the six continents. Each paper received at least two reviews, and authors were required to address review comments prior to presentation and publication.

Conducting CISSE 2009 online presented a number of unique advantages, as follows:

- All communications between the authors, reviewers, and conference organizing committee were done on line, which permitted a short six week period from the paper submission deadline to the beginning of the conference.
- PowerPoint presentations, final paper manuscripts were available to registrants for three weeks prior to the start of the conference
- The conference platform allowed live presentations by several presenters from different locations, with the audio, video and PowerPoint presentations transmitted to attendees throughout the internet, even on dial up connections. Attendees were able to ask both audio and written questions in a chat room format, and presenters could mark up their slides as they deem fit
- The live audio presentations were also recorded and distributed to participants along with the power points presentations and paper manuscripts within the conference DVD.

The conference organizers and we are confident that you will find the papers included in this book interesting and useful.

Khaled Elleithy
Tarek Sobh
Magued Iskander
Vikram Kapila
Mohammad Karim
Ausif Mahmood

Bridgeport, Connecticut
January 2010

Acknowledgements

The 2009 International Conferences on Telecommunications and Networking (TENE 09), Engineering Education, Instructional Technology, Assessment, and E-learning (EIAE 09), and Industrial Electronics, Technology & Automation (IETA 09) and the resulting proceedings could not have been organized without the assistance of a large number of individuals. TENE, EIAE, and IETA are part of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE). CISSE was founded by Professors Tarek Sobh and Khaled Elleithy in 2005, and they set up mechanisms that put it into action. Andrew Rosca wrote the software that allowed conference management and interaction between the authors and reviewers online. Mr. Tudor Rosca managed the online conference presentation system and was instrumental in ensuring that the event met the highest professional standards. We also want to acknowledge the roles played by Sarosh Patel and Ms. Susan Kristie, our technical and administrative support team.

The technical co-sponsorship provided by the Institute of Electrical and Electronics Engineers (IEEE) and the University of Bridgeport is gratefully appreciated. We would like to express our thanks to Prof. Toshio Fukuda, Chair of the International Advisory Committee and the members of the TENE, EIAE and IETA Technical Program Committees.

The excellent contributions of the authors made these world-class conference proceedings possible. Each paper received two to four reviews. The reviewers worked tirelessly under a tight schedule and their important work is gratefully appreciated. A complete list of reviewers is provided in the book.

Khaled Elleithy
Tarek Sobh
Magued Iskander
Vikram Kapila
Mohammad Karim
Ausif Mahmood

Bridgeport, Connecticut
January 2010

Contents

Reviewers List

Abbas	Riazi	Hiten	Choudhury	Peng	Su
Ahmad Sofian	Shminan	Ilias	Karasavvidis	Pinde	Chen
Alejandro	Regalado	Ilie	Borcosi	Popescu	Marian
Alen	Hajnal	Inci	Saricicek	Razib Hayat	Khan
Andrew	Dimmond	Indrajit	Mandal	Reza	Vahidnia
Anu	Gupta	Ioana	Chan Mow	Riyanto	Sigit
Ardian	Greca	Jan	Genci	Sadhana	Talele
Ausif	Mahmood	Joao	Ferreira	Sadia	Sultana
Barry	Stein	Jozef	Halko	Sat	Sharma
Biju	Issac	Jozef	Mašcenik	Seibu	Mary Jacob
Bingyang	Zhou	Kabaou	Mohamed Ouwais	Shafqat	Hameed
Bo	Yuan	Katarina	Jelemenska	Shen	Chao
Bojan	Medjo	Keith	Griffin	Smita	Shukla
Chi	Thai	Krystyna Maria	Noga	Sridhar	Chandran
Chris	Panagiotakopoulos	Liciniu-Alexandru	Kovacs	Stefan	Belev
Christian	Ayala Esquivel	Lucia A.	Fernandes	Sylvia	Encheva
David	Moore	Luy	Nguyen Tan	Teodor	Iliev
David	Wyld	Maduranga	Dassanayake	Teresa Cristina	Bessa Assunção
Denilson	Oliveira	Magued	Iskander	Thomas	Yang
Deniss	Kumlander	Margareth	Stoll	Vanessa	Cox
Dimitrios	Tomtsis	Martina	Hedvicakova	Vini	Kanvar
Dmitry	Zub	Massimo	Esposito	Xingguo	Xiong
Edson	Cavalcanti Neto	Mazeyanti	M Ariffin	Yenumula	Reddy
Enrique	Araiza Montes de Oca	Michel	Owayjan	Yoichiro	Iwasaki
Eric	Gamess	Mikhail	Stetsenko	Yuanyao	Liu
Ernest	Hibbs	Mofakharul	Islam	Yuval	Cohen
Fawzi	Albalooshi	Mohammad	Karim	Zheng (Jeremy)	Li
Grofu	Florin	Mohammad	Muqri	Zhenxing	Luo
Harrison	Yang	Monika	Verma		
Hector	Barbosa Leon	Ognjen	Kuljaca		

Education in the Clouds: How Colleges and Universities are Leveraging Cloud Computing

David C. Wyld and Rusty L. Juban
Southeastern Louisiana University
Department of Management – Box 10350
Hammond, LA 70402-0350 USA

***Abstract*- Cloud computing has been touted as a revolutionary concept in computing. In this article, we introduce the cloud computing concept, and then discus the implications for the use of the cloud model in higher education. Due to budget constraints and the power of computing "on-demand," many believe that colleges and universities will be at the forefront of the shift to using cloud-based resources. We examine the impact of cloud computing on both research and university computing operations. We see that globally, this shift is already making a difference in how universities procure and use IT resources. In conclusion, we advance a six-stage Cloud Migration Strategy for introducing and implementing cloud computing in institutions. We also discuss the necessary changes in policies and procurement, as well as security and reliability concerns for institutions implementing cloud computing.**

I. INTRODUCTION

A. *A Different Way of Computing*

I need a computer. Actually, I need the processing power of *hundreds* of computing hours. Heretofore, if I was a researcher running data or testing a model, that meant using solely the computing power available on my campus computing system. For a major operation, that might mean waiting behind other faculty and student projects, and then having to run my data over days at a time. Today, that computing power can be had at my fingertips in a matter of minutes - or even seconds.

Likewise, my email, my files, my programs were all formerly on my computer – or on my campus' mainframe. Today, those operations - and my data - may reside on servers in Washington State – or in Bangalore. And this may not just be for my computing needs. Today, it may be for my entire campus and all of the institution's students and faculty.

Welcome to the world of cloud computing!

B. *The Cloud Computing Concept*

The Economist [1] reminds us that: "Computing has constantly changed shape and location—mainly as a result of new technology, but often also because of shifts in demand" (n.p.). We have seen revolutionary computing technologies - truly "game changing" concepts - come about roughly once each decade in the "modern era" of computing since around 1945 when computing came to mean computations performed by a machine, not by man. From the mainframe era of the 1960s to the advent of minicomputers in the 1970s, the personal computer in the 1980s, the growth of the Internet and the Web in the 1990s, and the explosion of cell phones and other smart, Web-connected devices in the past 10 years (*see Figure I).*

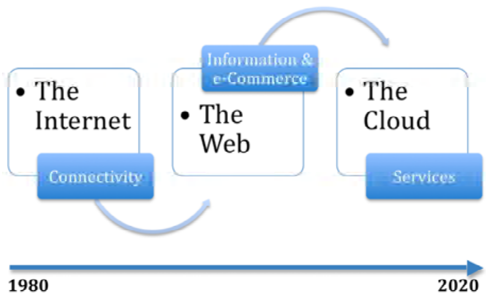


Fig. I. From the Internet to Cloud Computing.

Now, many think that cloud computing will be "the next big thing." Indeed, Gartner [2] believes that in the end, the impact of the cloud model will be "no less influential than e-business" (n.p.). If industry analysts are correct, we thus stand at an inflection point – a true paradigm change – in the evolution of computing.

The basic idea behind cloud computing is that *anything* that could be done in computing – whether on an individual PC or in a corporate data center – from storing data to communicating via email to collaborating on documents or crunching numbers on large data sets - can be shifted to the cloud. As can be seen in Table I, cloud computing encompasses a wide variety of offerings, including: SaaS (Software as a Service), PaaS (Platform as a Service), and IaaS (Infrastructure as a Service) [3].

Cloud computing has now become "shorthand" for the larger trend of computing services delivered over the Internet [4]. From the perspective of the market analyst, IDC [5], cloud computing represents "an emerging IT development, deployment and delivery model, enabling real-time delivery of products, services and solutions over the Internet." As Lohr [6] characterized it: "Cloud computing — in which vast stores of information and processing resources can be tapped from afar, over the Internet, using a personal computer, cell phone or other device — holds great promise…to cut the costs, complexity and headaches of technology for companies and government agencies."

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_1, © Springer Science+Business Media B.V. 2010

Table I
VARIANTS OF CLOUD COMPUTING

Level	Label	Description
User Level	SaaS "Software as a Service"	Companies host applications in the cloud that many users access through Internet connections. The service being sold or offered is a complete end-user application.
Developer Level	PaaS "Platform as a Service"	Developers can design, build, and test applications that run on the cloud provider's infrastructure and then deliver those applications to end-users from the provider's servers.
IT Level	IaaS "Infrastructure as a Service"	System administrators obtain general processing, storage, database management and other resources and applications through the network and pay only for gets used.

Source: Adapted from Rayport and Heyward [3].

Certainly, one of the hallmarks of cloud computing is that it enables users to interact with systems, data, and each other in a manner "that minimizes the necessary interaction with the underlying layers of the technology stack" [7]. According to the *Cloud Computing Manifesto*: "The key characteristics of the cloud are the ability to scale and provision computing power dynamically in a cost efficient way and the ability of the consumer (end user, organization or IT staff) to make the most of that power without having to manage the underlying complexity of the technology" [8].

The Economist [9] captured the meaning of this trend in stating: "The plethora of devices wirelessly connected to the Internet will speed up a shift that is already under way: from a 'device-centric' to an 'information-centric' world….(and) as wireless technology gets better and cheaper, more and more different kinds of objects will connect directly to the cloud." Technology guru Clay Shirky perhaps put it best when he said: "What is driving this shift is a change in perspective from seeing the computer as a box to seeing the computer as a door" [3]. The emerging cloud computing paradigm is thus based on a "user-centric interface that makes the cloud infrastructure supporting the applications transparent to the user" [10].

C. Overview

How does this new, on-demand, information-centric model of computing fit in the world of higher education – and what does it entail for research, for collaboration and for communication in colleges and universities? This article examines the early evidence from the field and discusses the practical and institutional implications. It concludes with a Cloud Migration Strategy for college and university IT executives to follow as they seek to best integrate cloud computing into their overall IT strategies.

II. Cloud Computing in Universities Today

For universities, migrating to cloud-based services affords them the ability to provide improved collaboration and research capabilities, while at the same time, providing an opportunity to cut IT costs while providing the same – or better – levels of computing services. Magnified by the need to pare overhead costs at a time when public and private institutions are grappling with significant budget shortfalls, cloud computing allows universities to not just use the resources of commercial cloud providers – many of which are available to them either for free or at reduced costs. With the cloud model, students and faculty can take advantage of the ability to work and communicate from anywhere and on any device using cloud-based applications.

The benefits for higher education center upon the scalability and the economics of cloud computing. These will be discussed in subsequent sections.

A. Scalability of Resources

One of the most important impacts of cloud computing will be the notion of computing power on-demand, in that: "When you radically democratize computing so that anyone has access at any moment to supercomputer-type capacity and all the data storage they need" [11]. This "democratization" of computing processing and storage power could have profound implications in everything from scientific inquiry (by making no problem too big to compute) to new enterprise formation (by drastically reducing the need for upfront investment in IT resources – and the people to support and maintain them) to public agencies (by making IT more affordable and available to governments at all levels and in all locales). Thus, we may be seeing a truly new era, where through democratizing computing technology, this will help to bring "the benefits of high-powered computers and communications to all" [12].

Cloud computing is a revolutionary concept in IT, due to an unprecedented elasticity of resources made possible by the cloud model. In everyday use, elasticity is commonly thought of not just as the ability of an object to stretch out when needed, but to also contract as necessary (think of a rubber band or a bungee cord). In computing terms, elasticity can be defined as: "The ability of a system to dynamically acquire or release compute resources on-demand" [7]. Under the cloud model, organizations that need more computing power have the ability to "scale-up" resources on-demand, without having to pay a premium for that ability. Say, for instance, that a researcher or a department has large, batch-oriented processing tasks. The individual or group can run the operations far faster than previously and at no additional costs, "since using 1000 servers for one hour costs no more than using one server for 1000 hours" [13]. This unique attribute of cloud computing is a commonly referred to as "cost associativity," and it allows for computational needs to be addressed far faster and far cheaper than in the past. In short, cloud computing gives organizations – even individual users – with unprecedented scalability. Graphically, this is depicted in Figure II (*The Cost Associativity Principle in Cloud Computing*).

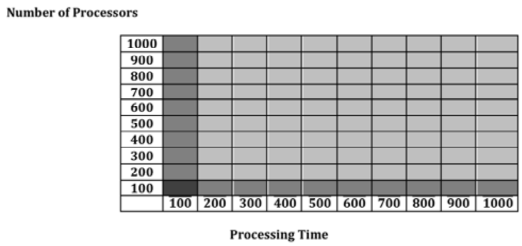


Fig. II. The Cost Associativity Principle in Cloud Computing
Source: Based on Armbrust, et. al. [13]; Reynolds and Best [14].

Additionally, where in the past only the largest universities have had supercomputing capabilities cloud computing, with number-crunching capabilities available on an on-demand basis, affords researchers anywhere to scale their computing power to match the scale of their research question - bringing supercomputing to the mainstream of research. As Delic and Walker [15] characterized it, cloud computing might just "enable new insights into challenging engineering, medical and social problems," as researchers will now have newfound capabilities "to tackle peta-scale type(s) of problems" and to "carry out mega-scale simulations." Craig A. Stewart, Associate Dean for Research Technologies at Indiana University, recently remarked that with cloud computing, "You reduce the barrier to use advanced computing facilities" [16].

We have seen the reduction of barriers already paying dividends in research. At pharmaceutical giant Eli Lilly, researchers needed to queue their projects to run in Lilly's internal data center. This process to provision enough server capacity for their respective projects often meant a delay of up to two months waiting on their data run. Today however, with cloud computing, research scientists can today provision the necessary processing capacity for their projects in five minutes. This allows researchers at Lilly and other research organizations to crunch data and test theories in ways that may have gone unexplored in the prior era where they would have been dependent solely on in-house computing resources [17]! Similar experiences are being reported at universities, both in the U.S. and abroad. For instance, at the International Institute of Information Technology in Hyderabad, India, Associate Professor Vasudeva Varma reports that the ability to run more data from more experiments more quickly has resulted in more publications for faculty in the Institute's Search and Information Extraction Lab, which he heads [18].

B. *Economics of Computing*

There is much discussion about the whole concept of "free" pricing for products and services today – and many of the email, storage, hosting, and applications that are at the forefront of cloud computing today are indeed free. The most notable of these are the product offerings of Google (Gmail, GoogleApps, GoogleDocs, and others). Much attention has been devoted to the concept of "freeconomics," most notably the recent book by *Wired* magazine editor Chris Anderson [19] entitled, *Free: The future of a radical price*. Most consumer-level cloud offerings would be labeled a "freemium," which is a free version that is supported by a paid, premium version. Such freemiums are becoming an emergent business model, as they are particularly popular among online service and software companies [19]. And, when faced with competing against "free" alternatives, older, more established companies have seen users migrate to the *gratis* alternative. Indeed, some see an entire "Culture of free" emerging, where from music to entertainment to news to software, people are coming to expect that free is the price they should pay [20].

In the corporate computing market, as software, hardware and processing power, and storage capacity become more and more commoditized, cloud computing becomes a free – or lower cost – alternative to the way things have been done for decades. As DiMaio [21] remarked: "Why should I bother looking for an email client to replace Outlook and coexist with my newly installed OpenOffice, if I can get email and office suite as a service with somebody like Google at a fraction of the cost and - most importantly - giving up the IT management burden too? Why are we talking about moving servers from Windows to Linux when the real question is why do we need to have our own servers in the first place?"

Already, there have been many campuses that have switched to Google or Microsoft-hosted email. Google and Microsoft host email for over four thousand colleges and universities, not just in the U.S., but in over 80 countries worldwide. In fact, almost half of all campuses are now making use of hosted email services [13]. The switch to hosted services is paying significant dividends for the early adopting institutions. By switching to Gmail, Notre Dame reports that it saved $1.5 million in storage and other tech costs, while at the same time, finding that their students' satisfaction with the campus' email rose by over a third! Likewise, institutions (such as Arizona State and Washington State) are consistently reporting at least six figure annual savings from switching to Google or Microsoft hosted systems [22]. Even more importantly, by switching to hosted email and productivity software, the job and focus of college IT staff can be changed. As Pepperdine University's CIO Timothy Chester observed, his smaller IT staff can be more productive. He recently stated: "We want our staff working more with students and faculty and less on the nuts and bolts of delivering technology" [22].

Certainly, as challenging budgetary times have been forecast to persist across higher education for the next few years – at least, there will likely be even greater pressures on colleges and universities to replace "paid" software and computing resources with "free" or low-cost cloud alternatives. From the cloud provider standpoint, Google has stated that its incentive in providing such free services to universities is to create "relationships for life" with students and faculty [23].

III. ANALYSIS

Many in higher education are coming to concur with the opinion of Keahey [24], who plainly stated: "I believe that this (cloud computing) will be the model of the future. This will absolutely drive cost savings for the universities." Across higher education, the cloud computing landscape should be quite active over the next few years, as we will see both coordinated efforts and "rogue" operations that will test how and where cloud computing can be effectively applied. As we have seen, colleges and universities will in many instances lead the way. These entities will continue to do so, based on their need for computing power on demand and for providing the types of ready – and in many cases free – IT resources – to their faculty and students. With pressure to reduce the fixed costs of higher education – and IT being a very rich target – the shift to cloud may be more forced in some cases than may be dictated by the on-the-ground circumstances. Indeed, some of the most exciting uses and best practices for cloud computing could well come from the world of higher education.

We have seen predictions that due to the cost and operational benefits of cloud computing, more and more companies will find themselves outsourcing most – if not all – of their IT to cloud providers, creating what has been termed as "serverless organizations" [25]. Indeed, it has been predicted that organizations of all sizes will find it beneficial to concentrate on and optimize their business processes by outsourcing the IT function. So, why not "serverless universities"? By outsourcing almost all of IT and all data storage/handling – this may be a viable proposition for colleges and universities, particularly as cloud offerings expand and are made more secure and reliable.

As we have seen in this article, there are certainly discussions and embryonic efforts underway – both in the U.S. and abroad - as public and private universities examine how to best made the cloud-concept work for they and their students and faculty. Universities are beginning to work collaboratively in the cloud to pool their IT resources. Already, this has occurred in Virginia and North Carolina. In the Commonwealth, a dozen colleges and universities have come together to form the Virginia Virtual Computing Lab [16]. Such efforts allow institutions to cut their IT costs by reducing their need for software licensing, for upgrade capabilities, and for perhaps maintaining their own data centers, all while improving the IT resources for their faculty and students. Already, by shifting to cloud offerings, North Carolina State University has been able to dramatically lower expenditures on software licenses and simultaneously, reduce the campus' IT staff from 15 to 3 full-time employees [18].

Additionally, there have been calls for the federal government to take the lead to create a cloud computing environment, to be available for use by all colleges and universities nationwide. In doing so, proponents argue for the economic and educational benefits that such a resource would provide, as it would democratize computing technology and "level the playing field" so all students and faculty could have access to the scale and type of computing power enjoyed only by elite institutions [26].

IV. CONCLUSION

A. A Cloud Migration Strategy for Higher Education

It is important to bear in mind the advice of Higginbotham [27] that "cloud computing is a tool, not a strategy." IT leaders in higher education will thus be well-advised to take a programmed, assessment of how cloud computing can fit into their overall IT strategy, in support of the mission and overall strategy of their institution. This should take a 6-step process, as shown in Figure III.

Fig. III. From the Internet to Cloud Computing.

The Cloud Migration Strategy begins with learning about the basics of cloud computing – through attending seminars, networking, talking with vendors, and reading articles such as this one. Given that cloud computing represents a new paradigm in computing technology, it will be important for technology transfer to occur – the "techies" in and outside of the institution will need to go the extra mile to educate and inform the "non-techie" amongst their ranks and constituencies as to the merits and value of cloud computing. It will be especially important to devote sufficient funding for research to establish how cloud computing is working – or not working – in various areas in the university and across institutions, so as to ground policies and develop best practices in regards to the use of cloud computing.

Then, IT executives should conduct an honest assessment of their institution's present IT needs, structure, and capacity utilization. In a cloud computing environment, where resources can be added – or subtracted – based on needs and demand, it will be critical for IT managers to *honestly* assess their institution's IT baseline for faculty, students and operations. In looking at data center utilization, it will be vital to look at what resources are used all the time and are necessary for day-to-day operations to establish a baseline for internally-hosted operations. Only then can one look at whether to continue to host "excess" capacity in the data center or to contract for cloud services as needed to scale-up to meet demands for greater amounts of computing resources.

University IT leaders should then pick one area – even one specific project - to "cloud pilot" and assess their ability to manage and bring such a project to fruition. As with any new technology, we are seeing a great deal of pure experimentation with cloud computing – "science project" like work for the most part up till now. All of us who use the Internet are experimenting with cloud applications in our daily lives – from Twittering to Gmail to using photo-sharing sites. In the same way, we are seeing organizations conducting cloud computing trials - what Woody [17] termed as "science experiments" in the use of the technology. Such efforts that are far away from their core IT operations and many times on (or trying to connect) the periphery of the organization. . Many times – even in the public sector and especially on campuses, these experiments may be "rogue" operations – taken on by individuals and units to test the utility of the technology. These are important efforts, and they should be supported – and reported within and outside the institution – so that others in the IT and the wider community can learn of the successes – and the downsides – of operating in the clouds. Thus, it will be vitally important to share both "best practices" and "lessons learned" in cloud computing. Indeed, many predict that such "science projects" in large and small organizations will drive the eventual acceptance and adoption of cloud computing [28].

After the internal assessment and external outreach stemming from the pilot effort, they should then conduct an overall IT cloud-readiness assessment to determine if they have data and applications that could readily move to a cloud environment and if a public/private/hybrid cloud would be suitable or useable for these purposes and rank-order potential projects. Finally, it is time to begin a cloud rollout strategy – gaining buy-in from both institutional leadership and IT staffers and communicating with both internal and external stakeholders as to the goals, progress, and costs/benefits of each cloud project. This is where the cloud goes from being a test effort to become more mainstream in the way the university manages its data, its operations and its people. It becomes part of "normal" operations, just as other prior tech innovations (from telephony to fax to the Internet to email and to social media) have become IT tools, used in support of the institution's IT strategy and more importantly, its overall *strategy*.

At this point, the process enters the final stage – call it "continuous cloud improvement" – to where the institution continues to move appropriate data and applications to the cloud – and perhaps even back from the cloud to internally-hosted operations, if necessary, based on a thorough and continuous assessment of the appropriate use of cloud technologies for their particular university.

B. *Implications for Higher Education*

The shift to more cloud-based applications will indeed bring newfound capabilities to communicate, collaborate and conduct research to university faculty, staff and students. However, it will also necessitate a flurry of policy decisions that will need to be made and operational rules that will need to be implemented. For instance, there will have to be IT policy decisions made as to who can access what files and what type of access they will have (i.e. read-only, editing access) [29]. The shift will also necessitate institutions to examine how cloud computing will secure and procure their computing environment.

Indeed, one of the principal concerns about cloud computing whether it is secure and reliable. Unfortunately, worries over cloud reliability and availability – or specifically, the lack thereof when such instances arise - are not just theoretical, as there have been well-publicized outages of many of the most popular public cloud services [30, 31]. And, as Schwartz [32] astutely pointed-out, when cloud service outages or inaccessibility occur, "most of the risk and blame if something goes wrong will fall directly on the shoulders of IT -- and not on the cloud computing service providers."

Security concerns may indeed impede the shift to cloud-based models. As with prior shifts in information technology with the advent of the Internet and the Web, the introduction of e-mail, and the explosion of social media, their growth and adoption rates have been slowed by initial fears – some justified and some very unjustified – over security concerns and the loss of control over data and operations [33]. Certainly, privacy and security questions will need to be addressed as institutional data and applications move into a cloud environment. Indeed, analogies have been drawn between the advent of cloud computing today with the introduction of wireless technologies a decade ago [34]. Finally, security is undoubtedly a hard metric to quantify. And, all too often, from the perspective of Golden [35] and other observers, the IT community has a somewhat damaging tendency to treating all risks – whatever the real nature of them – as the very worst case scenario and not judging the true impact – and likelihood – of their occurrence.

Finally, universities' often outdated and byzantine procurement rules and regulations, some of which may even preclude the use of cloud computing in select instances, will need to be changed to be more cloud-friendly and encourage the savings and efficiencies that can come from this new model of IT [36]. There will also need to be changes made in not just the language, but in the mindset of contracting for computing services. For while IT administrators look at capacity and systems, end users look to performance. As Jackson [37] recently put it, the key metric will now become: "When I sit down at that computer, do I see the functionality I need?"

In time, we may look back on the latter portion of this first decade of the new millennium as a true turning point in the history of computing. The transition however will take years, perhaps even decades, and "we're not going to wake up tomorrow and get all our computing requirements through a socket in the wall" [38]. However, all signs point to a true, campus-led revolution in computing.

REFERENCES

[1] *The Economist Editorial Staff*, "Let it rise," *The Economist*, October 23, 2008. [Online]. Available: http://www.economist.com/specialreports/displayStory.cfm?story_id=12411882 [Accessed: March 3, 2009].

[2] Gartner (2008), *Press Release: Gartner says cloud computing will be as influential as e-business*, June 26, 2008. [Online]. Available:

http://www.gartner.com/it/page.jsp?id=707508 [Accessed: February 24, 2009].
[3] J.F. Rayport and A. Heyward, *Envisioning the cloud: The next computing paradigm*, March 20, 2009. [Online]. Available: http://www.marketspaceadvisory.com/cloud/Envisioning-the-Cloud.pdf [Accessed: March 24, 2009].
[4] J. Hempel, "Salesforce hits its stride," *Fortune*, February 16, 2009. [Online]. Available: http://money.cnn.com/2009/02/16/technology/hempel_salesforce.fortune/index.htm?postversion=2009021710 [Accessed: March 8, 2009].
[5] F. Gens, "Defining 'cloud services' and 'cloud computing,'" *IDC Research*, September 23, 2008. [Online]. Available: http://blogs.idc.com/ie/?p=190 [Accessed: May 2, 2009].
[6] S. Lohr, "I.B.M. to help clients fight cost and complexity," *New York Times*, June 15, 2009. [Online]. Available: http://www.nytimes.com/2009/06/15/technology/business-computing/15blue.html?_r=1&emc=eta1 [Accessed: June 17, 2009].
[7] K. Langley, "Cloud computing: Get your head in the clouds," *Production Scale*, April 24, 2008. [Online]. Available: http://www.productionscale.com/home/2008/4/24/cloud-computing-get-your-head-in-the-clouds.html [Accessed: June 3, 2009].
[8] *The open cloud manifesto: A call to action for the worldwide cloud community* (Draft 1.0.9). *OpenCloudManifesto.org*, March 30, 2009. [Online]. Available: http://opencloudmanifesto.org/opencloudmanifesto1.htm [Accessed: April 1, 2009].
[9] *The Economist Editorial Staff*, "On the periphery," *The Economist*, October 23, 2008. [Online]. Available: http://www.economist.com/specialreports/displaystory.cfm?story_id=12411896 [Accessed: March 3, 2009].
[10] IBM, *White Paper - Seeding the Clouds: Key Infrastructure Elements for Cloud Computing*, February 2009. [Online]. Available: ftp://ftp.software.ibm.com/common/ssi/sa/wh/n/oiw03022usen/OIW03022USEN.PDF [Accessed: February 22, 2009].
[11] N.G. Carr, "The many ways cloud computing will disrupt IT," *InfoWorld*, March 25, 2009. [Online]. Available: http://www.tmcnet.com/usubmit/2009/03/25/4084363.htm [Accessed: April 1, 2009].
[12] J. Foley, "Report: Cloud computing could be bigger than the Web," *Information Week*, March 20, 2009. [Online]. Available: http://www.informationweek.com/cloud-computing/blog/archives/2009/03/google_report_c.html?catid=cloud-computing [Accessed: April 24, 2009].
[13] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica and M. Zaharia, *Above the clouds*, University of California, UC Berkeley Reliable Adaptive Distributed Systems Laboratory, February 2009. [Online]. Available: http://d1smfj0g31qzek.cloudfront.net/abovetheclouds.pdf [Accessed: March 10, 2009].
[14] E. Reynolds and C. Bess, "Clearing up the cloud," *Cutter IT Journal*, June/July 2009, vol. 22, no. 6/7, pp. 14-20.
[15] K.A. Delic and M.A. Walker, "Emergence of the academic computing clouds," *Ubiquity*, August 5, 2008, vol. 9, no. 31, pp. 2-5.
[16] J.R. Young, "3 ways web-based computing will change colleges," *The Chronicle of Higher Education*, October 24, 2008. [Online]. Available: http://chronicle.com/cgi2-bin/printable.cgi?article=http://chronicle.com/free/2008/10/5611n.htm [Accessed: June 5, 2009].
[17] T. Woody, "The future of cloud computing," *Portfolio*, March 9, 2009. [Online]. Available: http://www.portfolio.com/views/columns/dual-perspectives/2009/03/09/Todays-Weather-Report [Accessed: May 8, 2009].
[18] K. Ng, "Campus on a Cloud," *FutureGov*, August 2009, vol. 6, no. 4, pp. 34-38.
[19] C. Anderson, *Free: The future of a radical price*. New York: Hyperion.
[20] Knowledge@Wharton, "How about free?" University of Pennsylvania, March 4, 2009. [Online]. Available: http://knowledge.wharton.upenn.edu/article.cfm?articleid=2169 [Accessed: June 10, 2009].
[21] A. DiMaio, "Do vendors have a vision for the future of government?" *Gartner*, April 8, 2009. [Online]. Available: http://blogs.gartner.com/andrea_dimaio/2009/04/08/open-source-in-government-is-irrelevant/ [Accessed: July 18, 2009].
[22] J. Caplan, "Google and Microsoft: The battle over college e-mail," *Time*, August 14, 2009. [Online]. Available: http://www.time.com/time/printout/0,8816,1915112,00.html [Accessed: August 27, 2009].
[23] S. Coughlan, "Google's e-mail for universities," *BBC News*, June 11, 2007. [Online]. Available: http://news.bbc.co.uk/2/hi/uk_news/education/6741797.stm [Accessed: September 1, 2009].
[24] J. Staten, "Cloud maturity is accelerating," *Forrester.com*, July 10, 2008. [Online]. Available: http://blogs.forrester.com/it_infrastructure/2008/07/cloud-maturity.html [Accessed: February 28, 2009].
[25] Appirio, Inc., *Press Release: Top 10 cloud computing predictions for 2009*. December 18, 2008. [Online]. Available: http://www.appirio.com/about/pr_predictions-09_121808.php [Accessed: January 12, 2009].
[26] W. Hurley, "Higher education needs a national computing cloud," *InfoWorld*, January 25, 2009. [Online]. Available: http://www.infoworld.com/d/cloud-computing/higher-education-needs-national-computing-cloud-404?source=fssr [Accessed: June 1, 2009].
[27] S. Higginbotham, "Cloud computing is a tool, not a strategy," *GigaOm*, February 19, 2009. [Online]. Available: http://gigaom.com/2009/02/19/cloud-computing-is-a-tool-not-a-strategy/ [Accessed: March 11, 2009].
[28] J. King, "5 key questions about cloud storage," *Computerworld*, July 13, 2009. [Online]. Available: http://www.computerworld.com/s/article/340471/Cloud_Storage_Illuminated [Accessed: July 23, 2009].
[29] D. Beizer, "Cloud computing comes into focus," *Government Computer News*, June 11, 2008. [Online]. Available: http://gcn.com/articles/2008/06/11/cloud-computing-comes-into-focus.aspx [Accessed: February 7, 2009].
[30] S. Gaudin and D. Nystedt, "Gmail outages could turn off enterprises," *Computerworld*, September 10, 2009. [Online]. Available: http://www.computerworld.com/s/article/342708/Gmail_Outages_Could_Turn_Off_Enterprises?taxonomyId=16 [Accessed: September 12, 2009].
[31] C. Waxer, "Is the cloud reliable enough for your business?" *Computerworld*, June 1, 2009. [Online]. Available: http://www.computerworld.com/action/article.do?command=printArticleBasic&taxonomyName=Servers+and+Data+Center&articleId=339152&taxonomyId=154 [Accessed: June 15, 2009].
[32] E. Schwartz, "The dangers of cloud computing," *InfoWorld*, July 7, 2008. [Online]. Available: http://www.infoworld.com/article/08/07/07/28NF-cloud-computing-security_1.html. [Accessed: August 1, 2009].
[33] Kelton Research, *2009 Cloud Computing Survey*, February 2009. [Online]. Available: http://www.avanade.com/people/thought_detail.aspx?id=70 [Accessed: May 8, 2009].
[34] D. Beizer, "NIST creates cloud-computing team," *Federal Computer Week*, February 25, 2009. [Online]. Available: http://fcw.com/articles/2009/02/25/nist-cloud-computing.aspx [Accessed: July 2, 2009].
[35] B. Golden, "The case against cloud computing, part two," *CIO*, January 29, 2009. [Online]. Available: http://www.cio.com/article/478419/The_Case_Against_Cloud_Computing_Part_Two [Accessed: June 4, 2009].
[36] M. Weigelt, "Kundra aids search for procurement leader," *Federal Computer Week*, July 31, 2009. [Online]. Available: http://fcw.com/Articles/2009/08/03/WEEK-Kundra-aids-OFPP-search.aspx [Accessed: August 7, 2009].
[37] J. Jackson, "New metrics needed for cloud computing," *Government Computer News*, July 30, 2009. [Online]. Available: http://gcn.com/Articles/2009/07/30/New-metrics-for-cloud-computing.aspx?p=1 [Accessed: August 3, 2009].
[38] K. Melymuka, "The end of corporate IT," *Computerworld*, May 9, 2005. [Online]. Available: http://www.computerworld.com/managementtopics/management/story/0,10801,101573,00.html [Accessed: January 30, 2009].

Weak Orderings of Knowledge States

Sylvia Encheva and Sharil Tumin*
Stord/Haugesund University College, Faculty of Technology, Business and Maritime Sciences, Bjørnsonsg. 45, 5528 Haugesund, Norway
University of Bergen, It Dept, P.O. Box 7800, 5020 Bergen, Norway*
sbe@hsh.no, edpst@it.uib.no

Abstract - **This article discusses assessment of knowledge states of students enrolled in a new subject. The above-mentioned knowledge states belong to a single study unit. Applying methods from the theory of weak orderings assures also firm quantitative support for subsequent tuning of current course materials, hints and diagnostic tests.**

I. Introduction

Establishing the knowledge state of a student in a subject is crucial for providing him/her with individual help. Subsets of relevant examination questions and certain skills from a branch of knowledge are listed as examples of knowledge states in [12]. It is quite important to underline that not all possible subsets of such items turn out to be knowledge states.

Outcomes of diagnostic tests usually give an indication about the general status of students' knowledge. At the same time, they do not address the problem of classification of knowledge states. The latter is strongly related to identification of existing misconceptions and lack of adequate knowledge. This work presents an approach that can be used in the process of solving these problems.

Based on students' responses a content provider can follow the process of response clusterings and make appropriate adjustments in contents of a problem or the related materials or both. Response clusterings can be used to track down the effect of assistance provided to each student and thus contribute to a better individualized assistance.

The rest of the paper is organized as follows. Section 2 contains information about related work. Section 3 is devoted to ordered sets. Section 4 relates various personal responses and Section 5 is devoted to a system description. Section 6 contains the conclusion of this work.

II. Background

Navigational structures in an automated tutoring system should prevent students from becoming overwhelmed with information and losing track of where they are going, while permitting them to make the most of the facilities the system offers, [9]. In that framework which supports personalized learning a new topic is recommended to a student after assessing his/her knowledge. The system developed as a result of that framework responds to students needs according to their learning orientations.

Artificial intelligence techniques are widely used for facilitating a learning process tailored to individual needs. Among the most successful computer-based instructional tools, seriously relying on such techniques, are intelligent tutoring systems.

An intelligent tutoring system is defined in [25] as a "computer system" based on "artificial intelligence" designed to deliver content and to provide feedback to its users. In [26] an intelligent tutoring system is considered to be a software agent that provides customized and adaptable instructions and feedback to learners. The topic received a lot of attention from the research community since the 1980's. However, it seems that the process of providing adequate help to users based on their individual responses is still open.

A cognitive model represents the knowledge that an ideal student would possess about a particular subject. Developing a cognitive tutor involves creating a cognitive model of students' problem solving by writing production rules that characterize the variety of strategies and misconceptions students may acquire, [16, 17, and 18]. Cognitive tutors have been successful in raising students' math test scores in high school and middle-school classrooms, but their development has traditionally required considerable time and expertise, [8]. An investigation of whether a cognitive tutor can be made more effective can be found in [15].

The process of delivering hints in an intelligent tutoring system has been discussed in [24]. Taxonomy for automated hinting is developed in [20]. The role of hints in a Web based learning systems is considered in [7] and [8]. Usability issues of a Web-based assessment system are discussed in [13] and [21].

A model for detecting students' misuse of help in intelligent tutoring systems is presented in [3]. A proliferation of hint abuse (e.g., using hints to find answers rather than trying to understand) was found in [1] and [15]. However, evidence that when used appropriately, on-demand help can have a positive impact on learning was found in [19] and [23].

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_2,

III. Ordered Sets

Two very interesting problems about orderings are considered in [5], namely the problem of determining a consensus from a group of orderings and the problem of making statistically significant statements about ordering.

An ordered set (or partially ordered set or poset) is an ordered pair $(P,\leq)$ of a set P and a binary relation '$\leq$' contained in $P\times P$ called the order (or the partial order) on P such that '$\leq$' is reflexive, antisymmetric and transitive.

A relation I is an *indifference* relation when given AIB neither $A>B$ nor $A<B$ has place in the componentwise ordering. A partial ordering whose indifference relation is transitive is called a *weak ordering*.

Let ω*1*, ω*2,* ω*3* be weak orderings. The ω*2* is between ω*1* and ω*3* if each decision made by ω*2* is made either by ω*1* or ω*3* and any decision made by both ω*1* and ω*3* is made by ω*2*, i.e. $\omega 1 \cap \omega 3 \subseteq \omega 2 \subseteq \omega 1 \cap \omega 3$.

The distance is defined as

$$d(\omega 1, \omega 2) + d(\omega 2, \omega 3) = d(\omega 1, \omega 3).$$

The distance is a metric in the usual sense, it is invariant under permutation of alternatives, and the minimum positive distance is 1. If two weak orderings agree except for a set of alternatives that is an open interval in both, then the distance may be computed as if the alternatives in this set were the only objects being ranked.

If given two alternatives, a person is finally choosing only one. The natural extension to more than two elements is known as the 'majority rule' or the 'Condorcet Principle'. A relation $R(L_1,L_2,...,L_k)$ is constructed by saying that the pair $(a,b)\in R$ if (a,b) belongs to the majority of relations L_i. The following three linear orderings $a\ b\ c$, $b\ c\ a$, and $c\ a\ b$, leading to $R=\{(a,b),(b,c),(c,a)\}$ (three-way tie), illustrate the 'paradox of voting', [4].

The probability of the voting paradox for weak orderings is calculated analytically for the three-voter-three-alternative case. It appears that the probability obtained this way is considerably smaller than the corresponding case for linear orderings, [22].

A. Social Welfare Function

A 'social welfare function' maps k-tuples of the set of linear orderings of any $b\subset A$ to single linear orderings of B, where A is a set of at least three alternatives, [2].

A social welfare function acts on k-tuples representing either weak or linear orderings of k individuals for m alternatives. The representation of an individual's ordering can be thought of as a column vector in which an integer in position i represents the preference level the individual assigns to alternative i. Thus k individuals presented with m alternatives can be illustrated by k-tuples of orderings in a $m\times k$ matrix of integers of preference levels.

Theorem 1 along with several axioms, [5] provides an existing social welfare function.

Theorem 1 *There is a unique social welfare function satisfying Axioms 0-4. It is given by: i is preferred to j if and only if the sum of the row of A corresponding to j is larger than the sum of the row of A corresponding to i, and otherwise the social ordering is indifferent between i and j.*

Two elements a and b where $a\neq b$ and $a,b\in P$ are comparable if $a\leq b$ or $b\leq a$, and incomparable otherwise. If $\forall a,b$ where $a,b\in P$ are comparable, then P is a chain. If $\forall a,b$ where $a,b\in P$ are incomparable, then P is an antichain.

B. Knowledge spaces

Within the theory of knowledge spaces, [11], [12] a knowledge domain is described by a set of problems Q. Then the knowledge state of a student is a set $Q' \subseteq Q$ that the student can solve. A knowledge structure of a subject is a set of all knowledge states.

A knowledge space is a closure system, [11] where closure system on a finite set M is a set F of subsets of M such that

$$M\in F \quad and\ C,C^1\in F \Rightarrow C\cap C^1\in F$$

IV. Geometric Interpretation of Knowledge States

Initially we are trying to establish the status of students' knowledge in a topic with respect to three knowledge states called *a*, *b* and *c*. This is realized by considering students' answers to three questions in a multiple choice test that is a part of a Web-based system. Students' responses are saved in a database. Depending on whether a provided answer is correct, partially correct or incorrect, the student is suggested appropriate reading materials, hints or examples. Afterwards she can repeat the test. The consecutive test contains new questions but related to the same knowledge states. This allows the course/test designer to follow each student's progress within any knowledge state and to obtain a clear picture about the degrees to which these knowledge states are reached by a particular student, and by a group.

If for example one or more of the knowledge states repeatedly appear to be difficult to be reached by a single student or by a group then the corresponding course materials and the related questions, have to be redesigned.

A knowledge state is assumed to be reached by a student if the student provides a correct answer to the question presented to her/him. The question is of course related to that knowledge state. A knowledge state is assumed to be partially reached by a student if the student provides a partially correct answer to the presented to her/him question, and knowledge state is assumed not to be reached by a student if the provided answer is incorrect.

Using nested parentheses one can represent the partial ordering structure internally in a given application. Let the alphabets in this definition be '( ), *a*, *b*, *c*'. The use of parentheses signifies the position of ordering of a particular element '*a*', '*b*', or '*c*', where '( )' implies correct answer, '(( ))' implies partially correct answer, '((( )))' implies incorrect answer. Parentheses are not used for the elements in the center of the graph.

Note that the set {*a, b, c*} is neither a chain no an antichain, i.e. a pair of different elements can either be comparable or incomparable.

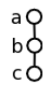


Fig. 1 Partial ordering $(a(b(c)))$

The ordering $(a(b(c)))$ in Fig.1 implies that knowledge state '*a*' is reached by that student, knowledge state '*b*' is partially reached, and knowledge state '*c*' is not reached by that student.

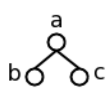


Fig. 2 Partial ordering $(a(bc))$

Fig. 2 illustrates a partial ordering $(a(bc))$ where knowledge state '*a*' is reached by that student, and both knowledge states '*b*' and '*c*' are partially reached.

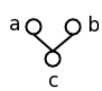


Fig. 3 Partial ordering $(ab(c))$

The ordering $(ab(c))$ in Fig. 3 implies that knowledge state '*c*' is partially reached by that student, while knowledge states '*a*' and '*b*' are reached.

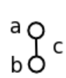


Fig. 4 Partial ordering $(a(b))c$

The ordering $(a(b))c$ in Fig. 4 implies that knowledge state '*a*' is reached by that student, knowledge '*b*' is partially reached, and the status of knowledge state '*c*' is unknown since the system has not registered any response from the student.

Suppose a particular study unit exhibits the following pattern of clustering (*ab*(*c*)) - 50%, (*a*(*bc*)) - 10%, and (*c*(*ab*)) - 20%. The rest of the 20% are distributed to other orderings and are therefore not significant. The significant factor is assumed to be 10% or above.

With this distribution, one can safely conclude that knowledge state '*a*' has been reached by the majority of students. Another interesting observation is that knowledge state '*c*' is reached by 20% of the students. Since the distance from (*ab*(*c*)) or (*a*(*bc*)) to (*c*(*ab*)) is the same, in this particular case, we can safely conclude that there are two types of students with different knowledge background in relation to the learning process of this particular study unit. Such observations are very useful for future tuning of course materials.

When a pattern of clustering does not show any significant value but is evenly distributed amongst the twelve ordering patterns (see f. ex. Fig. 5) then one can say that none of the provided materials was particularly helpful for the students' study progress. Both the course materials and the questions in diagnostic tests need to be redesigned.

Discussions on distances between different orderings are enclosed in the next subsections.

A. Orderings Where Two of the Knowledge States Either Reached or Partially Reached

The observed shortest distances between two orderings of knowledge states from a shown sequence are considered to be both either reached or partially reached.

1) From ordering *(a(bc))* to ordering *((ac)b)*

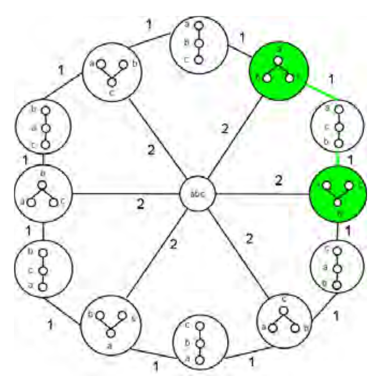


Fig. 5: Orderings *(a(bc))* and *(ac(b))*

In ordering *(a(bc))* we observe one knowledge state, '*a*', to be reached and two knowledge states, '*b*' and '*c*', to be partially reached, as shown in Fig 5. In ordering *((ac)b)* we observe two knowledge states, i.e. '*a*' and '*c*', to be reached and one knowledge state, '*b*', to be partially reached. Note that knowledge state, '*a*' is reached in both the orderings *(a(bc))* and *(ac(b))* while one of knowledge states, '*c*', is partially reached in ordering *(a(bc))* and reached in ordering *((ac)b)*. This indicates that the course materials, hints, questions, etc. related to knowledge states '*a*' and '*c*' do not currently need adjustments. The remaining question is how to improve on the provided assistance in a way that the knowledge state '*b*' is also reached. The distance between ordering *(a(bc))* and ordering *((ac)b)* is 2.

• From ordering *(ab(c))* to ordering *(c(ab))*

All knowledge states are reached in a different way in ordering *(c(ab))* in comparison to ordering *(ab(c))*. Thus knowledge states '*a*' and '*b*' being reached in ordering *(ab(c))* are partially reached in ordering *(c(ab))*, and knowledge state '*c*' is partially reached in ordering *(ab(c))* and reached in ordering *(c(ab))*. This indicates that the course materials, hints, questions, etc. related to knowledge state '*c*' do not currently need adjustments.

The remaining question is how to improve on the provided assistance in a way that knowledge states '*a*' and '*b*' remain reached. In similar occasions, it is recommended to pay particular attention to the contents of questions involved in establishing both orderings. The distance between orderings *(ab(c))* and *(c(ab))* is 4.

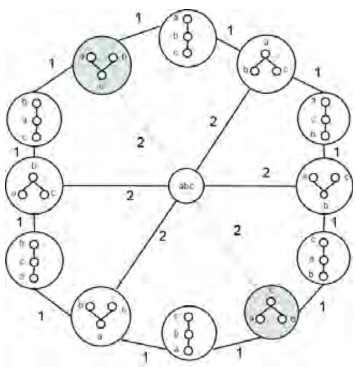


Fig. 6: Orderings *(ab(c))* and *(c(ab))*

• From ordering *(a(bc))* to ordering *(b(ac))*

Both orderings in Fig. 7 have the same structure. Two of the knowledge states, '*a*' and '*b*' are being respectively reached and partially reached in ordering *(a(bc))* and are partially reached and reached in ordering *(b(ac)))*, where knowledge state '*c*' is partially reached in both orderings

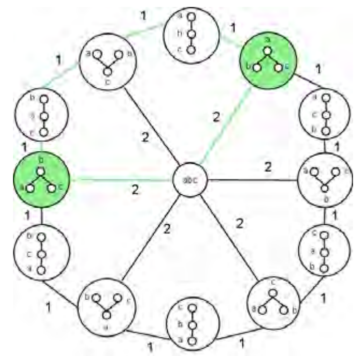


Fig. 7: Orderings *(a(bc))* and *(b(ac))*

This indicates that the course materials, hints, questions, etc. related to knowledge state '*b*' do not currently need adjustments. The work that has to be done in this case is related to knowledge states '*a*' and '*c*'. The fact that knowledge state '*a*' is only partially reached in ordering *(b(ac))* requires special attention to be paid to the involved questions and to the related course materials, hints, etc. Since knowledge state '*c*' remains partially reached in both orderings the course developer can concentrate on tuning the provided explanations, examples, hints, etc. The distance between orderings *(a(bc))* and *(b(ac))* is 4.

B. Orderings Where Knowledge States are Reached, Partially Reached or Not Reached

• From ordering *(a(c(b)))* to ordering *(c(a(b)))*

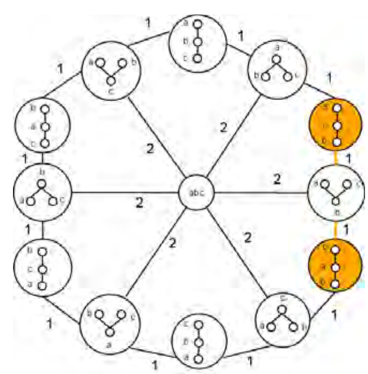


Fig. 8: Orderings *(a(c(b)))* and *(c(a(b)))*

Two of the knowledge states in Fig. 8, '*a*' and '*c*' are being respectively reached and partially reached in ordering *(a(c(b)))* and are partially reached and reached in ordering *(c(a(b))*, where knowledge state '*b*' is not reached in both orderings. This indicates that the course materials, hints, questions, etc. related to knowledge state '*c*' do not currently need adjustments. The work that has to be done in this case is related to knowledge states ‘a’ and ‘*b*’. The fact that knowledge state ‘*a*’ is only partially reached in the second ordering requires special attention to be paid to the involved questions and to the related course materials, hints, etc. Since knowledge state ‘*c*’ remains not reached in both orderings the course developer can work on the provided reading materials, explanations, examples, hints, etc. The distance between orderings *(a(c(b)))* and *(c(a(b)))* is 2.

• From ordering *(a(c(b)))* to ordering *(c(b(a)))*

All knowledge states change their status regarding the degrees to be reached, Fig 9. This occurrence requires very serious attention on the involved reading materials, explanations, examples, hints, questions etc. If it is due to a particular student response only then individual approach might be appropriate. The distance between orderings *(a(c(b)))* and *(c(b(a)))* is 4.

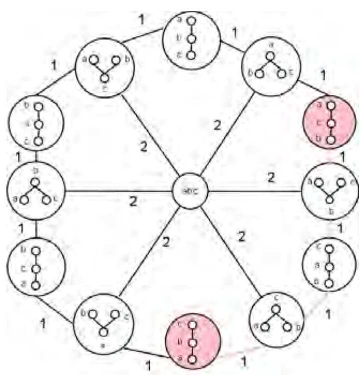


Fig. 9: Orderings *(a(c(b)))* and *(c(b(a)))*

• From ordering *(a(c(b)))* to ordering *(b(c(a)))*

Knowledge states ‘*a*’ and ‘*b*’ in Fig. 10 are being respectively reached and not reached in ordering *(a(c(b)))* and are partially reached and reached in ordering *(b(a(c)))*, where knowledge state ‘*c*’ is partially reached in both ordering. This indicates that the course materials, hints, questions, etc. related to knowledge states ‘*a*’ and ‘*b*’ have to be reconsidered. Since knowledge state ‘*c*’ remains partially reached in both orderings the course developer could concentrate on tuning the provided explanations, examples, hints, etc.

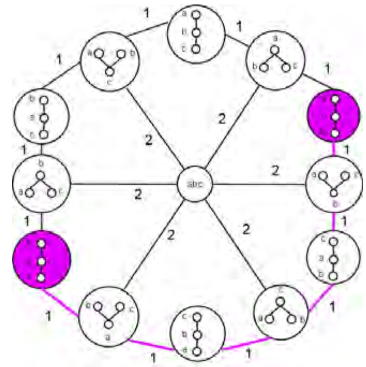


Figure 10: Orderings *(a(c(b)))* and *(b(c(a)))*

C. *Various Orderings*

Occurrences like those in Fig. 11 where one of the orderings is of type *(a(bc))* and the other one is an ordering like *(a(c(b)))*, *(c(a(b)))*, and *(c(b(a)))* are treated like the ones in the previous subsections. In an attempt to save space and keep to the geometrical interpretation, the couples of orderings to be considered are indicated with additional arrows instead of coloring the corresponding paths.

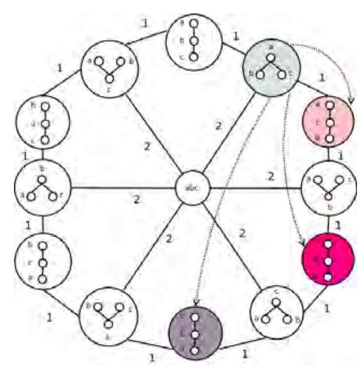


Fig. 11: Various orderings

Remark: With respect to orderings discussed in this section it is important to establish which group of students have a clear tendency to move within any of the ordering considered in subsections *A, B, C*. These groups are formed based on the amount of preliminary knowledge, current performance, learning styles, [14], learning orientations, [10] etc. Such data can be important while f. ex. updating instructions for future students and refining contents of questions in diagnostic tests.

V. System Description

A system prototype is build as a Web-based application using Apache HTTP server, mod_python module and SQLite database. The mod_python module provides programmable runtime support to the HTTP server using Python programming language. The application components are Web-based users interface, application logic and application interfaces written in Python, and a relational database. The system architecture is presented in Fig. 12.

The system contains also an interactive theoretical part (lecture notes), problems to be solved at home, tests assessing recall of facts, an ability to map the relationship between two items into a different context, and conceptual thinking, asking students to evaluate consequences and draw conclusions, and various quizzes.

The users, i.e. expert tutors, teachers, and students interact with the system using Web forms. Experts and teachers can submit and update data, while students can only view information.

The system provides recommendations on whether or not a student needs to take additional classes (courses) based on fuzzy dependencies.

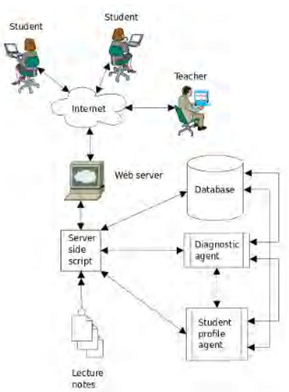


Fig. 12: System architecture

V. Conclusion

The work described in this article provides an approach that helps to locate significant gaps in students' knowledge. In addition, it gives an opportunity not only to follow the learning progress of individuals but also to recognize pattern clustering's within a study unit. The course developer can then easier find out which of the involved reading materials, hints and tests need further adjustments and what the future work should be concerned about.

References

[1] Aleven, V., and Koedinger, K. R.: Limitations of Student Control: Do Student Know when they need help? In G. Gauthier, C. Frasson, and K. VanLehn (Eds.), Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000 Berlin: Springer Verlag (2000) 292—303.

[2] Arrow, K. J. (1951). Social Choice and Individual Values. Wiley, New York, (1951).

[3] Baker, R.S., Corbett, A.T., and Koedinger, K.R.: Detecting student misuse of intelligent tutoring systems. Lecture Notes in Computer Science, Vol. 3220. Springer-Verlag, Berlin Heidelberg New Jork (2004) 531-540

[4] Berg, S.: Condorcet's Jury Theorem and the Reliability of Majority Voting, Group Decision and Negotiation, 5(3), (1996), 229-238

[5] Bogart, K. P.: Some social sciences applications of ordered sets. In: I. Rival, Editor, Ordered Sets, Reidel, Dordrecht (1982), 759–787

[6] Breuker, J., Components of Problem Solving and Types of Problems. In EKAW '94, (1994), 118-136

[7] Brunstein, A. and Krems, J.: Helps and Hints for Learning with Web Based Learning Systems: The Role of Instructions, Intelligent Tutoring Systems, Lecture Notes in Computer Science 3220, (2004), 794-796

[8] http://www.carnegielearning.com/web_docs/Whitepaper_Edweek.pdf

[9] http://ctat.pact.cs.cmu.edu/

[10] http://www.learning-styles-online.com/overview/

[11] Doignon, J.-P. and Falmagne, J.-C.: Knowledge Spaces. Springer-Verlag, 1999.

[12] Dowling C.E.: On the irredundant generation of knowledge spaces. Journal of Mathematical Psychology, 37, 1993, pp. 49-62

[13] Jan, N., Shun-Chieh Lin, Nancy P. Lin, Chang, C., A Trend-based Prediction System for Web User Behavior, WSEAS Transactions on Advances in Engineering Education, 5, 2008

[14] Keefe, J. W., Learning style: An overview. In NASSP's Student learning styles: Diagnosing and proscribing programs, (1979) 1-17. Reston, VA. National Association of Secondary School Principles.

[15] Koedinger K.R., McLaren B.M. and Roll I.: A help-seeking tutor agent. Proceedings of the Seventh International Conference on Intelligent Tutoring Systems, ITS 2004, Berlin, Springer-Verlag, (2004) 227-239

[16] Mayo, M. and Mitrovic, A.: Optimising ITS behaviour with Bayesian networks and decision theory. International Journal of Artificial Intelligence in Education 12 (2001) 124-153.

[17] Pecheanu E., Segal C. and Stefanescu D.: Content modeling in Intelligent Instructional Environment. Lecture Notes in Artificial Intelligence, Vol. 3190. Springer-Verlag, Berlin Heidelberg New Jork (2003) 1229-1234.

[18] Renkl, A.: Learning from worked-out examples: Instructional explanations supplement self- explanations. Learning and Instruction, 12, (2002) 529-556

[19] Stewart, K. L., and Felicetti, L. A., Learning styles of marketing majors. Educational Research Quarterly, 15(2), (1992) 15-23

[20] Tsovaltzi, D., Fiedler, A., and Horacek, H.: A Multi-dimensional Taxonomy for Automating Hinting, Intelligent Tutoring Systems, Lecture Notes in Computer Science 3220, (2004), 772-781

[21] Yeum, Y., Lee W., Usability Issues of Web-Based Assessment System for K-12 and Implementation of PAS System, WSEAS Transactions on Advances in Engineering Education, 9(4), (2007)

[22] Van Deemen, A., The Probability of the Paradox of Voting for Weak Preference Orderings. Social Choice and Welfare, 16(2), (1999)

[23] Wood, D.: Scaffolding, contingent tutoring, and computer-supported learning. International Journal of Artificial Intelligence in Education, 12 (2001) 280-292

[24] Zhou, Y., Freedman R., Glass, M., Michael, J. A., Allen A., Rovick, A. A., and Evens, M. W.: Delivering hints in a dialogue-based intelligent tutoring system, Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, (1999), 128-134

[25] http://design.test.olt.ubc.ca/Intelligent_Tutoring_System

[26] http://www.cs.ualberta.ca/ zaiane/courses/cmput605/2008-3.html

Problem Based Learning: Obtaining Enzyme Kinetics Parameters Integrating Linear Algebra, Computer Programming and Biochemistry Curriculum

Regalado-Méndez, A. [(a)], Báez-González, J., G. [(b)] Peralta-Reyes, E. [(a)], and Trujillo-Tapia, M., N. [(a)]
(a) Ingeniería Ambiental, Universidad del Mar-Puerto Ángel, Ciudad Universitaria S/N, Km. 1.5 Carretera Puerto Ángel–Zipolite, Puerto Ángel, Oaxaca, 70902, México.
(b) Departamento de Alimentos, Facultad de Ciencias Biológicas, Universidad Autónoma de Nuevo León, Av. Universidad S/N, Ciudad Universitaria San Nicolás de los Garza, Nuevo León, 66451, México.
E-mail: regalado@angel.umar

***Abstract*-This work is focused on a project that integrates the curriculum of biochemistry, lineal algebra and computer programming. The purpose is that students develop a software tool which calculates enzyme kinetic parameters based on proposed data. This program calculates such parameters using a linear regression of one of the linear forms of the Michaelis-Menten equations; moreover it characterizes the confidence of the lineal fit with the correlation coefficient. Once the different proposed steps were accomplished, we concluded that the purpose was satisfactorily reached with an increment in creative ability. The important is that the percentage of failure among students was 57%, 50%, 28% and 18% from 2005 to 2008, respectively.**

I. Introduction

In Mexico, problem based learning (PBL) has not been developed much, contrary to what is found in European countries and United States. In the last few years, this type of learning has become very popular [8], allowing the integration of different disciplines in to the curriculum. Frequently, students ask themselves about the usefulness of some courses, as they don't manage to see any concrete application. To ensure that students take a primary part in their own academic formation, the lecturer should assume a facilitator role, and encourage the students to motivate themselves in searching for possible solutions to a given problem, looking at it from a multidisciplinary viewpoint.

As shown in fig.1, students from the Universidad del Mar have high failure statistics which leads teachers to seek different pedagogic approaches. A project that allows integrating courses of biochemistry, linear algebra, and programming is proposed herein. It is intended for use with undergraduate students of the Universidad del Mar, Mexico. Until now, it has been applied to students in the Environmental Engineering program, but it can be expanded to students of natural science.

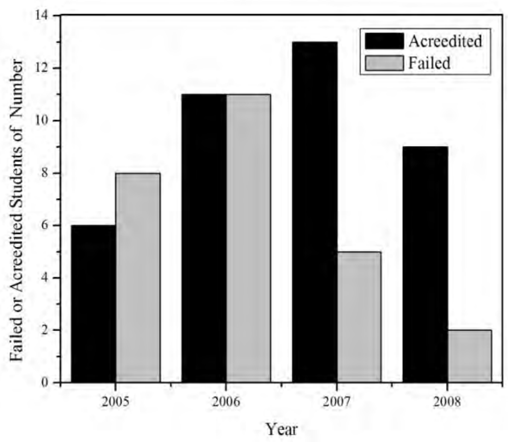


Fig. 1. Number of failed and accredited students each year.

To improve the Teaching-Learning process in Mexico, it is necessary that professors interconnect the knowledge obtained by the program curricula not only in a horizontal way but also vertically, with the objective to improve the cognitive processes of students in the first semesters, in order to form more critical and less procedural students [9].

The PBL structure was kept similar to the one adopted by Offman and Cadet, 2002, with 3 work sessions of 2-3 hours, the first with an introductory purpose. In these sessions, discussion is encouraged with a minimum involvement of the professor (facilitator) [1, 3, 6 and 7]. Each session is spaced by three days; in each a written report that answers specific questions is expected.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_3, © Springer Science+Business Media B.V. 2010

II. Problem Description

The PBL problem is: Determine the enzyme α-chymoptrypsin kinetic parameters, developing a software program. The rates of hydrolysis of N-acetyl-phenylalaninamide by α- chymoptrypsin at different concentrations are as follows [2]:

TABLE 1
EXPERIMENTAL DATA

[S] (mM)	V (mM de Alanine minute^{-1})
10.5	3.33×10^{-2}
15.1	4.55×10^{-2}
20.0	5.56×10^{-2}
35.0	7.15×10^{-2}
93.4	10.0×10^{-2}

A. *Part I (Session 1)*

The facilitator, in this section, should motivate the students to investigate subjects related with enzyme kinetics. In order to assist, he can propose the following questions: What are enzymes?

What are they useful for? How do they work? What is their reaction mechanism? Do models that allow predicting their reaction rate exist? If so, which would be the simplest?

B. *Part II (Session 2)*

Once the enzymatic kinetics is understood, the facilitator must introduce the mathematical tool and outline the following questions: What is the mathematical form of the models encountered? How can we turn the simplest model into a linear form? How can we determine the fit parameters, mathematically, in a linear equation?

C. *Part III (Session 3)*

In this section we will apply the programming theory, using flux diagrams to facilitate the program design. To develop software that will quickly obtain the kinetic parameters for any case study, using the linear equation of part II, the program must:

- Be able to read the data directly from keyboard capture or those present in a file with extension .dat.
- Calculate of kinetic parameters.
- Present them on screen and keep save as txt extension file.
- Characterize the quality of the adjustment realized.

III. Student Self-Assessment

In the self-assessment the students are asked to identify their skills and weaknesses in problems solutions. Students answer each of the following questions. When he/she finishes, he/she should grade themselves points between 0 and 10.

A. *Attitude*

1. If I didn't understand how to the begin problem. I tried understanding it. YES NO
2. I was honest with my self-learning (I didn't copy; I did my work without help). YES NO
3. I was responsible in my self-learning (I searched for information, I always looked more in-depth, I didn't stop with the information acquired, and I didn't depend on others to do my work). YES NO
4. I did all the indicated activities. YES NO
5. I actively participated in my team and contributed some individual ideas. YES NO
6. I was respectful in team participation (Taken decisions, agreements, and etcetera). YES NO

B. *Knowledge*

7. I understand all the concepts. YES NO
8. Now I could solve the problem the by myself. YES NO
9. From this example, I understand the importance of teamwork. YES NO

IV. Expected Answers

A. *Part I (Session 1)*

Enzymes are proteins that are highly specialized, due to the fact that they catalyze numerous reactions in biological systems.

In general, an enzyme gives the adequate environment for a determined reaction to be, energetically, more favorable. The particularity of an enzymatic catalyzed reaction is that it occurs in a specific place of the enzyme, the active site. The molecule that is fixed on the active site and on the enzyme is called substrate. The enzyme-substrate complex formed is of vital importance in order to define the kinetic behavior of the different catalyzed reactions. A simple enzymatic reaction can be represented by the Michaelis-Menten equation.

B. *Michaelis-Menten Equation*

Generally, it is accepted that the catalysis of enzymatic reactions needs the formation of a complex between the enzyme *E* and the substrate *S*. This enzyme-substrate complex *ES*, can dissociate to reform *E* and *S*, or can undergo some reorganization and form *E* and products *P*. The simplest mechanism of all can be represented as [2]:

$$E + S \underset{k_2}{\overset{k_1}{\rightleftarrows}} ES \underset{k_4}{\overset{k_3}{\rightleftarrows}} E + P \qquad (1)$$

When concentration *ES* is constant, the rate at which it is formed is the same as the rate of its decomposition. Therefore, according to the law of mass, the equilibrium state is:

$$k_1[E][S] + k_4[E][P] = k_2[ES] + k_3[ES] \qquad (2)$$

At the beginning of the reaction we have $[P] << << [S]$ and $[S] \approx$ cte, with the velocity V of P of formation as $V = k_3[ES]$, moreover the totality of the enzyme (E_{Tot}) can be complexes with the substrate or not, with $[E_{Tot}] = [E] + [ES]$, and the maximum rate V_{max} is reached when all the enzymes present are involved in the active enzyme-substrate complex, i.e. $V_{max} = k_3[E_{Tot}]$. Substituting the previous considerations in (2) and reorganizing the Michaelis-Menten equation can be obtained and its description is in the table 2:

$$V = \frac{[S]V_{max}}{[S] + k_m} \qquad (3)$$

TABLE 2
VARIABLE DESCRIPTION

Variable	Description
[E]	Enzyme Concentration
[S]	Substrate Concentration
[ES]	Enzyme-Substrate Complex Concentration (mM)
k_i	Reaction Kinetic constant
K_m	Constant for a determined enzyme and a given substrate (mM)
V	Velocity of product formation (mM min^{-1})
V_{max}	Maximum velocity of product formation (mM min^{-1})

C. Part II (Session 2)

As equation (3) is hyperbolic and obtaining the $V_{\max}$ and k_{m} parameters directly is complex, a first step is to make the Michaelis-Menten equation linear, as follows:

$$p(x) = a_0 + a_1 x \tag{4}$$

Equation (3) can be presented in the next lineal form, which is:

$$\underbrace{\frac{1}{V}}_{f(x)} = \underbrace{\frac{k_m}{V_{\max}}}_{a_1}\underbrace{\frac{1}{[S]}}_{x} + \underbrace{\frac{1}{V_{\max}}}_{a_o} \tag{5}$$

According to the case, one or the other equation can be adopted. The choice is basically guided by the experimental data that, once in the linear form have to be far enough from each other in order to obtain a confident linear curve. The following deviations to linearity can be related to impure substrate or enzyme, to the fact that more than one enzyme acts on the substrate, or to different types of inhibition of the system [4].

D. Determination of enzymatic parameters using minimum squares

Once the Michaelis-Menten linear equation is used, the minimum squares method is applied, which tries to apply the best curve upon experimental data. The method involves the sum of squares of the calculated distances between the ordinate at an experimental abscissa (x) of the line are polynomial function P (x_i) and the real function f(x) of the experimental data is minimum, i.e. [5]:

$$\sum_{i=1}^{m}\left[p(x_i) - f(x_i)\right]^2 = \sum_{i=1}^{m} d_i^2 \tag{6}$$

Substituting (4) in (6), we observe that there is a family of linear curves that pass through the points, nevertheless obtaining the minimum of (6) using the derivate of the function respecting to each of the variables (a_o and a_1) and setting to zero, a system of linear equations of the following form is generated:

$$\begin{aligned} m a_o + \sum_{i=1}^{m} x_i a_1 &= \sum_{i=1}^{m} f(x_i) \\ \sum_{i=1}^{m} x_i a_o + a_1 \sum_{i=1}^{m} x_i^2 &= \sum_{i=1}^{m} f(x_i) x_i \end{aligned} \tag{7}$$

Equations (7) can be resolved with the Cramer rule. Parameters (a_o and a_1) are given as:

$$\begin{aligned} a_o &= \frac{\left[\sum_{i=1}^{m} f(x_i)\right]\left[\sum_{i=1}^{m} x_i^2\right] - \left[\sum_{i=1}^{m} x_i\right]\left[\sum_{i=1}^{m} f(x_i) x_i\right]}{m\sum_{i=1}^{m} x_i^2 - \left[\sum_{i=1}^{m} x_i\right]^2} \\ a_1 &= \frac{m\sum_{i=1}^{m} f(x_i) x_i - \left[\sum_{i=1}^{m} f(x_i)\right]\left[\sum_{i=1}^{m} x_i\right]}{m\sum_{i=1}^{m} x_i^2 - \left[\sum_{i=1}^{m} x_i\right]^2} \end{aligned} \tag{8}$$

To obtain $V_{\max}$ and k_{m} parameters, from (5) it comes:

$$V_{\max} = \frac{1}{a_o} \quad and \quad k_m = \frac{a_1}{a_o} \tag{9}$$

To determine how confident the adjustment of minimum squares is, the correlation coefficient can be calculated (r^2) as:

$$r^2 = \frac{m\left(\sum_{i=1}^{m} x_i y_i\right) - \left(\sum_{i=1}^{m} x_i\right)\left(\sum_{i=1}^{m} y_i\right)}{\sqrt{\left[m\sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m} x_i\right)^2\right]}\sqrt{\left[m\sum_{i=1}^{m} y_i^2 - \left(\sum_{i=1}^{m} y_i\right)^2\right]}} \tag{10}$$

Depending of the value of this correlation coefficient, it is accepted that:

1)	*Perfect*	$r^2 = 1.0$	
2)	*Excellent*	$0.9 \le r^2 < 1.0$	
3)	*Good*	$0.8 \le r^2 < 0.9$	(11)
4)	*Medium*	$0.5 \le r^2 < 0.8$	
5)	*Bad*	$r^2 < 0.5$	

E. Part III (Session 3)

In this section, kinetic parameters are calculated.

F. Program flux diagram

Before doing the program, a blocks diagram will be designed, which will help to have a clear idea of the data needed, of the following steps, and of the desired results. The flux diagram is written in order to be translated in any programming language. An example is given in fig. 2.

Note: this section has greater complexity and the students develop the cognitive process in accordance to the mathematical sequential logic obtained in his/her academic trajectory. In this step, the student constructs new knowledge.

G. Example of developed software

Once the flux diagram is obtained in order to calculate the kinetic parameters of Michaelis-Menten equation, it is translated in MATLAB 7.6 language. An example of what can be achieved by students is given, in the format of an executable file, in order to be able to run the program without the use of MATLAB 7.6, as shown in fig. 2.

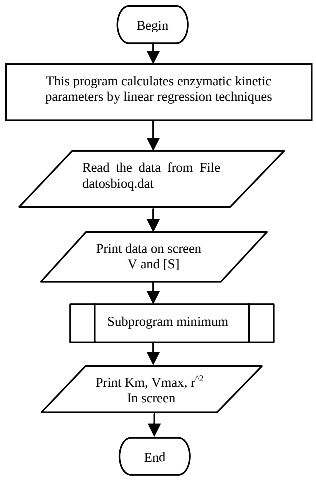


Fig. 2. Example of flux diagram.

If students want to read experimental data, they must check that the data are correctly written in the required emplacement, with the right filename (example datosbioq.dat).

To calculate the kinetic parameters we ran the CISSE2009.m program from MATLAB 7.6. A window opens that explains what the program does, as observed in fig. 3.

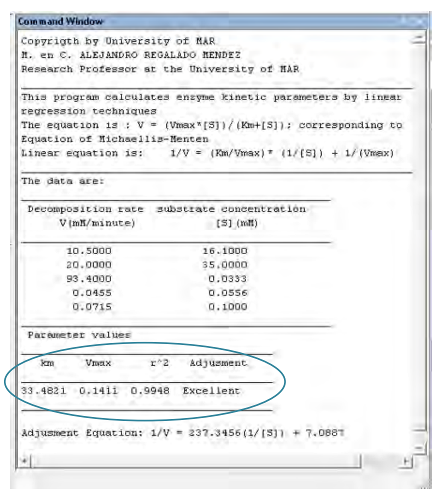


Fig. 3. Enzymatic kinetic parameters obtained.

In fig. 4, the linear adjustment and the experimental data are shown, characterizing the adjustment obtained.

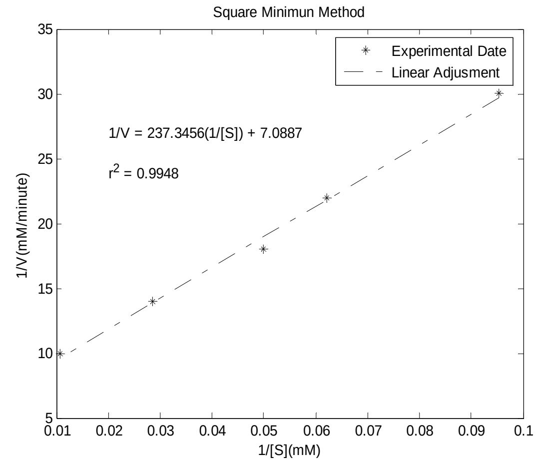


Fig. 4. Linear adjustment of Michaelis-Menten equation.

In addition to the parameter values, the correlation coefficient is given; in the study case, according to (11) it is excellent.

In Fig. 5, velocity degradation profile and experimental data with adjustment parameters are shown.

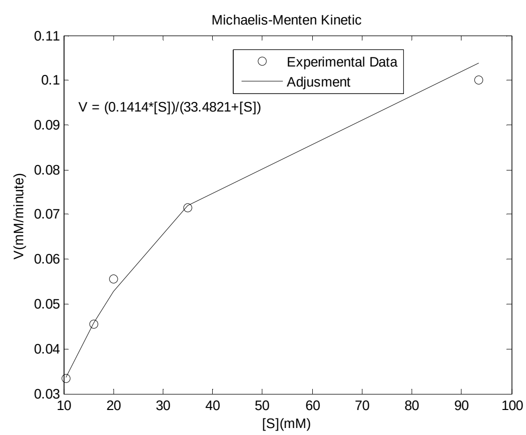


Fig. 5. Degradation rate profile of Michaelis-Meten equation.

The grade of student's understanding was greater within the proposed methodology, given that students participated highly, showing strong motivation while using programming as a tool to find the solution of a given problem integrating several curriculum courses. In Fig. 6 presents a picture of students on this activity.

Fig. 6. Working students.

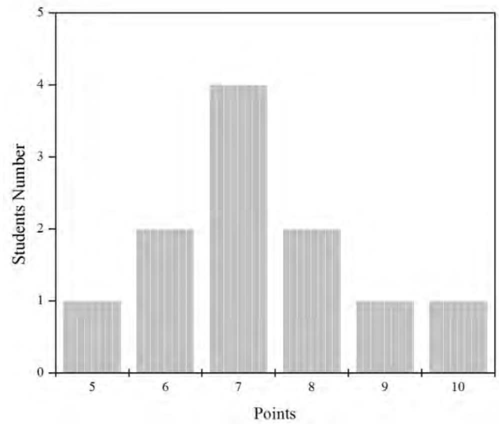


Fig. 8. Self-Assessment point.

The most difficult activity for learners was to establish the flux diagram, as it was in this part where they have to use sequential mathematic logics.

In Fig. 7. we have an analysis of the answers to each question in the student's self-assessment. Most students had a good attitude and understand the concepts.

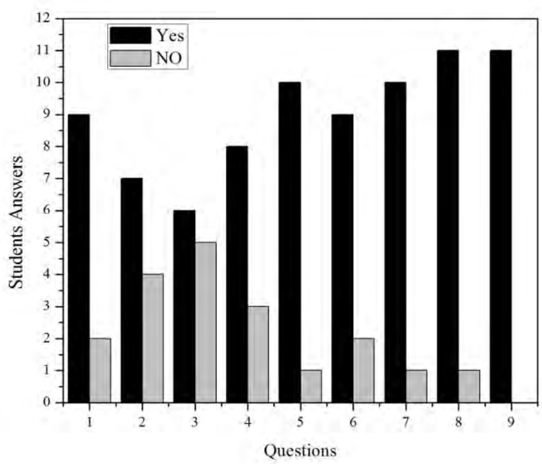


Fig. 7. Self-Assessment statistics.

In Fig. 8. shown, the self-assessment points of each student show that only one student self-reported that he failed, because he likes the traditional learning method more and most were comfortable with PBL, as demonstrated by above average points.

The computer programming was a pretext for students' motivation and they acquired their own responsibility knowledge to build their own success.

The methodology proposed considerably reduced the student's failure from 2005 to 2008, as shown in Fig. 9. The percentage of failure in 2005, 2006, 2007 and 2008 was of 57%, 50%, 28% and 18% respectively. Therefore PBL is a better way to give class.

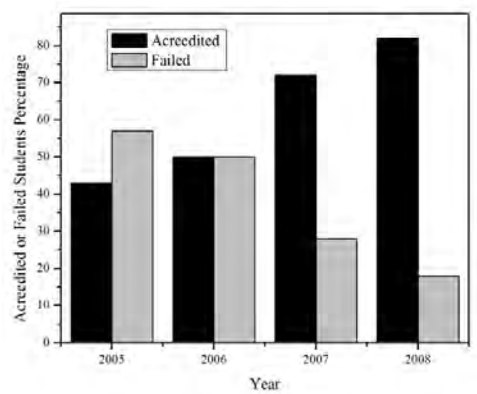


Fig. 9. Percentage of failed and accredited students each year.

I wish to mention that it is very important to listen to the students, because they can help to improve the learning process between professor and students.

V. Student Feedback

The students stated that, PBL is preferred, because:

i. They can acquire skills at their own pace within a problem solving environment.
ii. Learning takes place in a cooperative student-centered, environment.
iii. It opens their eyes to how university life should be.

iv. PBL improved their generic skills.
v. It makes us feel more comfortable to work in a group to solve problems.
vi. They said that "It really makes us work for ours studies and we feel intelligent doing it."
vii. We feel like engineers when we are doing the group assignments.
viii. The given problem excites the students' curiosity to know more.
ix. It can be described as a double edged sword, because we are required to do more than with traditional learning methods education.

VI. CONCLUSIONS

- Once the students concluded each step of the work program, the facilitator of this activity notes a cognitive improvement in the global understanding of the involved discipline, and the objective is satisfactorily attained.
- The PBL method is a strong tool to develop the creativity of students, as the prime objective of universities is to produce critical students.
- If the objective was attained for the students which had the PBL to solve, it is worth noting that the degree of understanding depends as well of the solidity of the knowledge background and the abilities acquired during the academic curriculum, and of the availability of specialists and the leadership of the facilitator involved.
- The use of PBL methodology can make the difference in the preparation and formation of students, and this is why it remains important that professors assume their educational role. This supports an updated commitment to the search for learning alternatives, as well as a responsibility of institutions to invest in teaching courses for the improvement of the educational processes.

ACKNOWLEDGMENT

The authors wish to thank to colleagues Kathleen Meredith Brown, Kathleen Dooley, Nina Carol Buck and Aitor Aizpuru for theirs revisions.

REFERENCES

[1] Dean, R., L., "Laboratory Exercises: Kinetic Studies with Alkaline Phosphatase in the Presence and Absence of Inhibitors and Divalent Cations", *Biochemistry and Molecular Biology Education*, 2002, Vol. 30 No. 6, pp. 4021-407.
[2] Hemberg T. "Abscising II as an inhibitor of α-quimoptripsinae", *Act Chem. Scand.*, 1967, Vol. 21, pp. 1666.
[3] Magued, I., Innovate Techniques in Instruction Technology, E-learning, E-assessment and Education, 2008, pp. 602.
[4] Montgomery R. and Swenson C. A., "Problemas Cuantitativos de las Ciencias Bioquímicas", Acribia, España, 1975, pp. 179-186.
[5] Nieves H. A. and Domínguez S. F., "Métodos Numéricos Aplicados a la Ingeniería", Continental, México, D.F., 2003, pp. 362-364.
[6] Offmann, B. and Cadet, F., "Problem-based Learning: Redox-active Disulfides in a Plant Light Switch", *Biochemistry and Molecular Biology Education*, 2002, Vol. 30 No. 4, pp. 249-254.
[7] Prausnitz J. M., "Hacia el fomento de la creatividad de los estudiantes" Educación en Química, 1997, Vol. 5 No. 3, pp. 156-159.
[8] Savin-Baden, M., "Problem-based Learning: Disciplinary Differences or Modes of Curriculum Practice? Who promised to deliver what in problem-based learning", *Biochemistry and Molecular Biology Education*, 2003, Vol. 31 No. 5, pp. 338-343.
[9] White H., B., "Problem-based Learning: Commentary Problem-based Learning and Becoming a Physician", *Biochemistry and Molecular Biology Education*, 2003, Vol. 31 No. 5, pp. 337.

Adaptive Assessments using Open Standards

Héctor G. Barbosa León, Francisco J. García Peñalvo, María José Rodriguez-Conde
University of Salamanca, Science Faculty, Plaza de los Caidos s/n, Salamanca, Spain 37008, Technological Institute of Colima, Av. Tecnológico s/n, Villa de Álvarez, Colima, México 28976
barbosah@usal.es, fgarcia@usal.es, mjrconde@usal.es

***Abstract* - This article presents the conclusions of a research workpreparedin the University of Salamanca, Spain, including the investigation, definition, construction and test of a model to define adaptive assessments considering the user´s presentation needs and also the traditional adaptation process using the level of complexity for every item. We applied innovatively some common open standards and specifications.**

I. INTRODUCTION

Students has been adopted the Internet as their first source of information. The Web and its intrinsic characteristics allow the content to evolve from a static to a dynamic organization by including adaptive processes, making the user´s experience to be improved and enhanced [1]. This content, developed in form of Units of Learning (UoL), include one or several learning objects and their related multimedia material.

But we must ensure that those developments have characteristics like be standardized, compatible and interchangeable, incrementing the characteristic of actual Web applications that are cross-platform systems.To achieve those characteristics, developers defined the Semantic Web, in which the content is seen in a semantic view instead of been seen in a syntactical one by using meta-data [2].

Among the huge amount of information in the Internet we will focus ourselves in instructional and educational institutions which have been using the Internet to incorporate the information and communication technologies in the learning and teaching processes in order to increase the quality, efficiency and dissemination of education.

We want to emphasize the role of the assessment activity inside the e-learning process, and how it can help to improve it to all participants: students, teachers and content designers.

II. IMPORTANCE OF THE ASSESSMENT ACTIVITY IN THE EDUCATIVE CONTENT

Traditionally, assessment activity has been seen like task aside of the e-learning process and there is a danger in focusing research on assessment specifically [3], as this tends to isolate assessment from teaching and learning in general.

For us, the assessment activity after the student took the educative lesson, could be seen as an element that closes and complete a circular activity, being an integral part of the learning process as well [4]. In addition, the assessment activity inside the e-learning process could be used to adapt the system by setting a new user knowledge level, evaluate, and setting new learning profiles, assign user grades and, in consequence, performing user content re-adaptation [1]. According to the Australian Flexible Learning Framework [3]assessment, especially when is included within a real learning task or exercises, could be an essential part of the learning experience, giving to the entire Web site the characteristic to adapt itself to the needs of the users.

Conceptualizing the learning process to its basic elements, we can identify at last the following elements:

- The educational material to be taught by the teacher in a classroom.
- The teaching and learning activities that take place in a classroom.
- The assessment activity to measure the student learning and,
- The report of the score results given by the teachers to the students. This conception is well suitable for the traditional educative process.

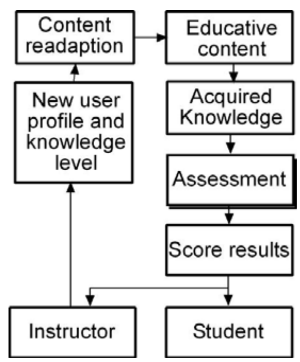


FIGURE 1. IMPORTANCE OF ASSESSMENT IN THE E-LEARNING PROCESS

In figure I, we characterize the importance of assessment in the learning flow. As we can note, the tests and evaluations not only are an integral part of the learning process, but also is an element that complete and close a circular activity, contributing as a feedback source for: the users (giving the scores and feedback), for the instructors (by giving support and feedback) and for the instructional designer (to update the contents of the learning system) as well. This circular conceptualization of the learning process allow us to see the significance of the assessment because it helps to the adaptation of the system by setting a new user knowledge level, evaluating and determining new learning profiles, assign the user grades and, in consequence, performing user content

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_4, © Springer Science+Business Media B.V. 2010

re adaptation. This is how we see the importance of the assessment task for the adaptation process.

Conclusively, we would like to mention an assessment categorization:

- Formative assessment: this is an assessment that helps to give a convenient feedback and motivation to the student and do not have scores. Also brings convenient feedback to designers of materials.
- Summative assessment: this is a scored assessment and gives a result to measure outcomes and success of the learner.
- Norm assessment: use the achievements of a group to set the standards for specific grades and is used in most universities.
- Criterion assessment: establish the standard (criterion) and mark students against it.

There is another kind of assessment called alternative assessment. Here, the integration of the assessment activity with learning processes and real-life performance as opposed to display of "inert knowledge". Know as authentic assessment it is very much based on the *constructivist* approach that enables students to demonstrate knowledge by working on authentic task, putting students undercontrol of their own learning [3].

III. CHARACTERISTICS FOR INTERNET-BASED SYSTEMS

Nowadays, learning systems in the Internet must be flexible and efficient, and one way to accomplish that is to be an open and standardized system. The following standards give to those systems some of the mentioned characteristics:

- The LTS (Learning Technology Standard) specification is a group of agreements about the characteristics that a learning element should have. The use of standards ensure instructional technologies to work with other systems (interoperability), follow-up information about learners and contents (manageability), generate objects that are usable in other contexts (reusability) and avoid obsolescence (durability) [5].
- IMS Specifications[1] are part of the LTS, developed in 1997 as a project of the National Learning Infrastructure Initiative at Educause. Its mission is to promote distributed learning environments. For the IMS many areas require interoperability when learning is distributed, thus it details a set of specifications that constitute a framework to interchange educational elements [6].
- Inside IMS specification one with special interest for our investigation is the IMS QTI (Question and Test Interoperability), for sharing test and/or items between compatible LMS. It defines a data model for the presentation of presentation of questions and test and the correspondent results.

[1] http://www.imsglobal.org

Those specifications allow the creation of many types of educational designs following a consistent notation to develop a system that could have a homogeneous implementation in several courses or learning contexts, giving the systems extra characteristics like compatibility and exchangeability. In addition, those specifications define an abstract data model using XML, supporting the deployment of item banks and well documented content format.

The potential benefit of Computer Based Test (CBT) are that a test item can attach various multimedia, such as graphic, audio or video, and reuse the test item easily, mostly if those are designed and developed with open specifications and standards, allowing to different information from test result from students can be collected and recorded by computers effortlessly and immediately in electronic portfolios (e-portfolios).

But most of CBT are linear that makes all students take the same number of test items in a specified order during a testing section. Nevertheless, this approach seems to imitate the paper-and-pencil (P&P) test because it merely presents the test in a digital format.

IV. CHARACTERISTICS OF ADAPTIVE SYSTEMS

In contrast to the linear test with fixed test items, Computer Adaptive Test (CAT), based on Item Response Theory (IRT) the presentation of each test item and the determination to finish a test are dynamically adopted based on the student´s answers.

One aim is to make those systems to work in adaptive learning systems and, given the fact that the assessment activity is an important and integral part of the e-learning process; it is desirable that this evaluation activity could be adaptable as well. If we want the assessment to be interoperable, compatible, and shareable, we must have to develop a standardized tool [8].

The implementation of the adaptability in the assessment activity could be performed by giving the instructor:

- Tools to design the assessment with several types of resources (text, graphics, video, animations) to improve the student´ understanding in the examination
- Design different types of assessment
- Define content structures and groups of students
- Manage the assessments and the questions linked to each assessment
- Define the grades for each question and assessment
- Define and manage the schedule for the assessment that will be taken.

The purpose of our work was to adapt the characteristics of the questions that are included in one individual's exam to their learning style [4],[5]. To do this, the IMS QTI specification allows the author to link the multimedia resources that could fit for three different user needs: visual, auditive or kinesthetic. That categorization fits with one of the learning styles definition[2].

[2] http:///www.ldpride.net/learningstyles.MI.html

Our aim was to define the primary resource (that is, the material presented to the user by default) and the equivalent resource (presented to the user taking into account his/her learning styles, physical disabilities, or preferences). For users with disabilities the user interface must be capable to make a transformation of the display using flexible controls. This adaptation is performed by matching the characteristics defined in the assessment item meta-data with those defined in the user profile in the Learning Management System.

Adaptation is another way to adjust to the user needs. This process is performed when the student is answering the questions, following the next sequence [9]:

- At the authoring stage, all items are categorized to estimate their level of difficulty.
- When the user starts the exam, he/she is presented with a pretest question to estimate his/her ability.
- The "best" next item is administered and the student responds.
- A new ability level is computed based on the responses of the student.
- Steps 3 and 4 are repeated until a stopping criterion is met.

Questions would be given randomly to eliminate cheating. The new ability level is computed based on equations described in the figure 1, used in the MANIC [10] system during the quiz sessions to compute the new estimated ability level.

The equation (1) is for correctly answered questions and the equation (2) is for wrongly answered questions.

NewValue = OldValue+OldValue(level of Question / 10)   (1)*
NewValue = OldValue+OldValue(5-level of Question / 10)   (2)*

V. ADAPTATION PROCESSES

I. Adaptation in the final presentation to the user

The learning object could adapt their final presentation taking into account the needs or preferences of the users. For this, the learning object containing an adaptable exam includes content for each learning style through the use of a specific environment. The LMS access the appropriate environment that fit to the user learning style or preference, showing the multimedia material [9], [11], [12],[13]. (Fig. 2).

II. Adaptation in the complexity level

Another process of adaptation is the level of complexity of the questions that are presented to the student (Fig. 2). The questions are selected by their level of complexity, taking into account the response of the student to the last question answered. If he/she answer correctly then the next question is of the same or higher complexity, if not, then the next question is of a lower complexity. This is the traditional adaptation process used by some developments in this area[3]. Please recall that the level of complexity of each item is given by the item' author in the definition stage, accordingly to his/her personal criterion.

One of the main characteristics of the IMS LD (Learning Design) specification is its potential to define adaptive characteristics covering the student preferences, prior knowledge and/or learning needs. To do this it is necessary to use this specification in the level B to define individual interactions because in this level we can use some elements like <*properties*> and <*conditions*>.

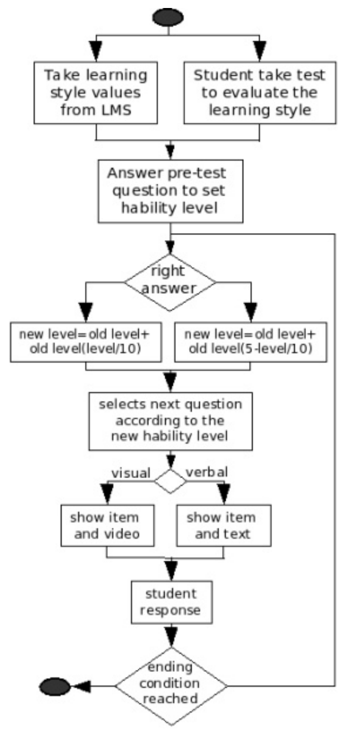


FIGURE 2 . ADAPTATION PROCESSES

The learning style values could be set from values stored in the user model in the LMS and stored in property elements (<*locpers-property*>, <*globpers-property*) to perform the adaptation. Finally, the <on-completion> element could be used to set the actions that will be done once certain action is performed. To perform adaptability to the learning style the LMS access the environment according to the need or selection of the user. From this, the questions set are presented to the student applying the adaptation algorithm proposed by [14].

Adaptation rule RUL1 is to perform adaptability according to the ability level and RUL2 is to adapt the presentation to the user learning style.

```
RUL1= IF <student>::(response,true)
    THEN
newLevel = oldLevel+
   oldLevel(oldLevel/10)
```

[3] http://www.imsglobal.org/cp

```
      ELSE
newLevel = oldLevel+
   oldLevel(5-oldLevel/10).
RUL2= IF <student>::(LS_visual)
     THEN
show item (newLevel,visual)
      ELSE
show item (newLevel,verbal).
```

VI. PACKAGE CONFIGURATIONS

The IMS CP (Content Packaging) specification is used when is necessary to transfer learning objects (lessons, exams or other material) between several Learning Management Systems. In this case we can use this specification to package and transfer assessment items between LMS. In the case of a simple element, the package will contain (a) the manifest (XML file called imsmanifest.xml), (b) the item (QTI XML format file) and, (c) any auxiliary files required by the item.

The IMS QTI and IMS LD (Learning Design) specifications allow their integration to define a learning object to evaluate the knowledge acquired by the students when interact with a unit of learning (UoL). The IMS LD includes activities of course instruction, modules, lessons and exams in a generic context (those considered as formative assessment) to support the recent knowledge or to give immediate feedback to the student.

The main structure defined in a IMS LD object is the manifest, containing the organization and resources. Inside the organizations section some elements are described, like the roles, properties, activities, environments and methods. The integration of the specifications could be done defining tags and instructions in the *imsld:properties* to control the visibility and order of the elements and the *imsld:conditions* to define decision structures.

The *environments* section is a container for the environment elements, each could be used to describe assessment items for a particular learning style. These structures (the environment ones) could be executed by the LMS in parallel, allowing multiple students of different learning styles to access their own adaptable elements (or an adaptable exam).

We propose three levels of integration [11] , going from simple single items to complete exams with adaptability characteristics.

A. First level of integration

In this level, the *testConstructor* could integrate a single item in a package containing the item description and the auxiliary files or the reference to them. The aim for this package is to be exported to the LMS, to be part of complete exams compatible with the IMS QTI specification or to be referenced in learning objects using IMS LD labels.

B. Second level of integration

In this level the *testConstructor* could define a group of single items to construct a complete exam, for a specific learning style (or, in consequence, for the generic learning style), selecting the single items from the database that could evaluate a specific unit of learning. In this level we include IMS LD labels in the *imsld:methods* section to define sequencing rules to show the items according to their level of complexity.

C. Third level of integration

In this level the *testConstructor* could construct a package containing a complete exam for each learning style. This package could be seen as a learning object to supply an exam for a specific learning style, including adaptation rules to show the items according the level of complexity. To define and construct the items and test, we consider using open specifications and standards to ensure the use of the products in a broader number of LMS.

The test object is divided in three main *testPart*(s), one for each kind of supporting material: audio, video or text. Inside each *testPart* there will be four assessment sections [13]:

- *assessmentSectionfirstNormalItem*: in this section the level of difficulty that will be presented to the student is set. The idea is to present a normal weight item to the student, if he/she responds correctly to it, then a heavier item will be presented to the user, if he/she responds right again, then a heavy level is set.
- *assessmentSectionlowWeight*: include the item set for questions with light weight.
- *assessmentSectionnormalWeight*: contain the item set for questions with normal weight.
- *assessmentSectionhighWeight*: enclose the item set for questions with heavy weight.
- *assessmentSectionsetDifficulty*: contains preconditions and branch instructions to guide the flow to the right *assessmentSection*, in concordance to the complexity level established.

This approach to organize individual assessment items into complete exams allows us to include adaptive metadata in a form of precondition and branching instructions. The precondition directives govern the access to test sections or to a single item according to a given condition, and the branching commands allow to the LMS to select the next section, test or individual item based on the fulfillment of a condition. We want to delineate this information to construct the adaptable navigation taking into account the weight or item difficulty.

VII. ADAPTIVE TEST

An adaptive test is a learning object organized in such way it can be used for compatible LMS to give each student with an adapted test using the parameters stored in the user model. The authoring process allows the definition and construction of different products [11]:

- As separate single items that could be referenced by compatible external tools using the IMS QTI specification.
- As complete test that could be used in a session by LMS as a part of a learning activity.
- As a learning object referenced from a learning unit defined using the IMS LD specification in which some conditions are defined to select the *testPart* using the value stored in the user model.

The elements to perform the adaptation are:

- User preference: stored in the user model with values of “visual”, “audio” or “text.
- Result of the first normal-level question stored in the *unifiedResult* variable, with values of “Y” or “N”. This variable is special because look to unify the result for the different kind of questions (simple choice, multiple choice, etc.).
- The result stored in the following questions in the variable *unifiedResult*, with values “Y” and “N”.

In the authoring process the author select from the list, those items that will be included in the test. The tool shows the items grouped by the multimedia files referenced by them. Once the selection is made, the tool construct the XML-QTI file with the *testPart*, *assessmentSection* sections and the reference to the external QTI files for each question and references to the external multimedia files as well. At the end, a file with the IMS CP instruction is generated and packed as .zip format; this file is the one that will be used by the LMS.

Also in the authoring process is possible to define the following kind of test:

- Sequential test: This test does not include branch rules, presenting the questions in the sequential order they were selected from the list.
- Random test: This test does not include branch rules, but the ordering label is set to shuffle.
- Adapted test: This test includes the adaptation rules with the *branchRule* and *preCondition* labels to control the items presented to the users.

VIII. ADASAT TOOL

We develop an authoring tool to construct items and test with adaptation characteristics, following open standards and specifications like the IMS QTI, CP and LD and the XML language. The aim is to allow to the author to define items and test using a graphical interface starting from the definition of single questions to the construction of complete test packaged using IMS CP and delivered in a .zip package [13], [14].

The application covers the following activities: (1) Question Construction management, (2) Test Construction management, (3) User management, (4) Test management.

IX. RESULTS OBTAINED

To verify that an adapted test is better for the students in terms to ease their answer and understanding, we apply an exam to evaluate the level of knowledge in English language to a group of students in the University of Salamanca. We split this group into two: for the first group we applied the test with questions and accompanying multimedia material that matches their preferences of presentation (audio, text or video); for the second group we applied a test that did not match these preferences. Also, we applied a test a test to determine the learning style of each student, prior to the adaptive test.

Please, recall that this test also perform an adaptation in the level of complexity of each item, for this we organize the items grouping them according this characteristic and, in a higher level, we set a classification according to the multimedia material they are refereeing to.

The aim of this activity is to evaluate our thesis hypothesis that, if a student is presented with an adaptive test that matches its preferences of presentation of the accompanying multimedia material he/she could average better results in a test.

The evaluation included the following statistical test:

- Final note: To check if the final note of the test is higher for the students with adapted test.
- Time to response: To check if the total time to answer the test is lower when an adapted test is presented to the student.
- Level of satisfaction: To check if the level of satisfaction reported by the student is higher when the test is adapted to his/her preferences.

For the final note test the results were: the average final note obtained by the students with an adapted test was 6.6 points, whilst for students without adapted test, the final note was of 5.9 points (t test = 0,836).

In the time of response the results were: the average time for an adapted test was 17 minutes whilst for an not adapted test was 20 minutes (t test=0.964).

The level of satisfaction was obtained from the students themselves from a poll after they answer the test. It cover level of satisfaction from 1 (lower satisfaction) to 5 (highest satisfaction). The results were: for an adapted test 3.9 and for not adapted test 3.6 (t test=0.821).

Bring to mind that, the statistical test was applied only to verify the results for the adaptation process in the presentation of the multimedia material, but this test also performs the adaptation in the complexity level for all participants.

Also, we think that these results must be taken with some considerations at the time to make generalizations. In other words the results obtained in this work may change in other learning environment and stated that this is an internal work. Upon this, we consider that our conclusions are a science contribution for future works in this area.

X. CONCLUSIONS

Online assessment is an important step inside the e-learning process, helping to improve the learning and teaching experience.

According to the new developments in the area of e-learning, we can see that most of these developments look to be compliant with accepted standards like the IMS QTI. This gives the convenience to those works to be interoperable and adaptable to several learning platforms. In concordance, referring to the assessment activity, we can think that it must be interoperable as well, because it is one element of the e-learning process and play an important role inside this task.

Adaptability is another key factor in assessment. Given the fact that assessment is an important element of the e-learning process and the fact that this process look to be interoperable, then we can think that the assessment tool could be used with different educative content administrators with different conceptualizations and ways to design an ways of purposes to the students. To face this situation it is necessary to develop a test with different types of resources, different kind of assessments, groups of students, kind of questions, etc.

Under this conceptualization, we created and applied an adaptive test to a group of students to evaluate the hypothesis that, if an assessment tool is adapted to the user`s needs or preferences we can reduce the time to answer the test while we increase the level of satisfaction of the user at the same time.

ACKNOWLEDGMENT

We want to thank to the group KEOPS (ref. TSI2005-00960) of the University of Salamanca for their contributions and ideas for the development of this work. In addition, we want to thank to the Education and Science Ministry of Spain, National Program in Technologies and Services for the Information Society, Ref.: TSI2005- 00960.

HéctorBarbosa thanks the National System of Technological Institutes (SNIT–Mexico) for its financial support.

REFERENCES

[1] Barbosa, H., Garcia F. Importance of the online assessment in the elearning process. 6th International Conference on Information Technology-based Higher Education and Training ITHET & IEEE, Santo Domingo.

[2] Barbosa, H., García, F., Adaptive Assessment Tool for the Hypertext Composer", ICTE2005. In proceedings of the 3th.Internatonal Conference on Multimedia and Information and Communication Technologies in Education: m-ICTE 2005, Cáceres, Spain, Vol.l 1, pp. 16-20.

[3] Booth, R., Berwyn, C. The development of quality online assessment in vocational education and training.Australian Flexible Learning Framework, Vol.1, 2003, 17.

[4] Barbosa, H., Garcìa, F.,An Authoring Tool to Develop Adaptive Assessments.In proceedings of 2006 Webist Congress, Setubal, Protugal, 2006.

[5] García, F., Berlanga, A., Moreno, M, N., Carabias, J. "HyCo and authoring tool to Create Semantic Learning Objects for Web-based E-Learning Systems".In Web Engineering Fourth International Conference, ICWE 2004. Published in N.Koch, P. Fraternali, M. Wirsing (Eds.). Series Lecture Notes in Computer Science. Springer Verlag. Vol. LNCS 3140, pp. 344-348.

[6] Berlanga A., Garcia, F., "Towards Semantic Metadata for Learning Elements". In proceedings of the fifth. International Conference on Information Technology Based Higher Education and Training ITHET 2004. Y. Akpinar (Ed.) (2004) pp. 572-577.

[7] Barbosa, H., Garcia, F. "Setting and Sharing Adaptive Assessment Assets". In proceedings of the SIIE congres, Leon Spain, 2006.

[8] Barbosa, H., García, F. "An Authoring Tool for Adaptative Assessment Items". In proceedings of IAIRA Congress. Bucharest, Romania, 2006.

[9] Barbosa, H., Garcia, F. "An Authoring Tool to Construct Adaptive Assessments". In proceedings of the 2006 EIAE Conference.Online, 2006.

[10] Mia,k., Stern, Woolf, B. "Curriculum sequencing in a Web-based tutor". Lecture Notes in Computer Science, 1452, 1998.

[11] Barbosa, H., Garcia, F., Rodriguez, M. "Defining Adaptive Assessments".In proceedings of the Webist Conference, Barcelona, Spain, 2007.

[12] Barbosa, H., Garcia, F. "Constructing Learning Objects for Adaptive Assessments". In proceedings of IASTED conference, Chamonix, France, 2007.

[13] Barbosa, H., Garcia, F., Rodriguez, M. "A Tool for Online Assessment in Adaptive e-Learning Platform".In proceedings of the eUniverSAL conference. Virtual, Salamanca, Spain. 2008.

[14] Barbosa, H., Garcia F., Rodriguez, M. "Construction of Assessments with Double Adaptation Processes". In proceedings of EIAE Congress, Virtual, 2008.

Online Academic Advising Support

Fawzi Albalooshi[1] and Safwan Shatnawi[2]
College of Applied Studies
University of Bahrain
Sakhir Campus, Bahrain Kingdom.
[1]fawzi@itc.uob.bh, [2]sshatnawi@asc.uob.bh

***Abstract*- The paper addresses academic advising – an important issue that is often not given enough attention by students and advisors alike. An online advising system is presented that can be utilized by students, advisors, and course timetable planners. Students are given informative advice on which courses to register for in the next semester and are informed of their remaining graduation requirements; advisors are able to see students' progress towards their graduation requirements; and timetable planners are given statistics on courses and sections requirements for the coming semester.**

***Index Terms*- Academic Advising, Automated Advising, Course Scheduling, Proactive Advising**

1. Introduction

Student advising is one of the important challenges facing academics, but direct consultation between instructors and students is another problem. Many students do not take the time and effort to see their advisors to plan their timetable before registration resulting in many registration issues and long queues for advising at registration time. Another problem facing academic institutes is planning ahead for the courses to be offered for the coming semester and determining the number of sections for each course. This difficulty is due to the fact that exam results for the courses in which students are registered in the current semester are not available until exams are over at the end of the semester. Determining students' graduation status is another time consuming effort for academics worldwide.

In this paper we present a system that accesses students' transcript records and generates a report for each student indicating the courses in which he must register in the coming semester based on rules explained later in the paper. Such details can only be exact towards the end of the semester once grades are input into the registration system. During the course of the semester, the proposed system can generate the students' requirements for courses including the number of sections required for the coming semester. This will enable the timetable committee to decide on the courses and sections to be offered and prepare the most suitable timetable for the students. Once grades are entered into the system at the end of the semester, a more accurate list of courses and sections can be generated before the start of the new semester. The proposed system helps students in advising, graduation, and timetabling. The students' advising and graduation support system is explained in Section 4 and the timetabling support system is explained in Section 5.

2. Literature Survey

Although the system described in this paper was primarily motivated by our need to optimize advising, to determine the courses to be offered and to utilize existing information available in the university registration system, there is no doubt that academics worldwide agree that proper advising is an important factor for students' successful progress in higher education. Many studies have been conducted to confirm this matter. Examples of such studies are reported in [1] and [3]. As a result, many academic institutes investigated the use of computer technologies in academic advising to overcome the difficulties experienced with traditional methods of advising. A sample list is as follows:

Reference [1] present a systematic view identifying key subsystems grouped under basic study planning and high level advising that can be automated to the benefit of advisers and advisees. Under basic study planning, the authors suggest that a number of basic queries be available such as course availability, courses in which a student can register, degree completion requirements, degree credit transfer possibilities, and time constraints. Reference [2] reports a system that offers advisees up-to-date online advising and related information including a recommended list of courses in which a student must register in the next semester in order to complete his degree requirements. The system has a web-based main page through which system users such as students, faculty, and administrative staff are allowed access to their respective sites. Reference [3] presents the motivations to develop FROSH [4] the automated advisor for freshman. Reference [5] presents a collaborative course recommender system that recommends elective modules to students based on the core modules they have selected. A "More Like This" recommender offers students similar modules to their first choice of electives in cases where no places are available or where timetable clashes occur. Reference [6] presents InVEStA, an expert system for advising students on the courses in which they must register next. The recommender system also generates a semester schedule based on the courses offered for the semester and the student curricula. Reference [7] presents a web-based interactive student advising system for course plan selection and registration using Java framework.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_5, © Springer Science+Business Media B.V. 2010

The works cited in this section are a selection of many systems developed by academic institutes for their specific programs that utilize the power of computer technologies to make student advising easier, accurate, and available to all.

3. Registration Data Description

Our university registration system maintains students' records, their transcripts, and courses details. Students' personal details include name, identification number, major etc. Students' transcript details include details of courses a student has completed, the grade obtained in each course, semester GPA and completed credit hours, and cumulative GPA and credit hours. All registration information just described can be accessed as an HTML format that cannot be handled directly. Web Semantic and text processing is necessary to extract the required information for our next processing phase. Tables 1 and 2 show student information extracted from the registration system using our text processing approach.

TABLE 1
STUDENT INFORMATION TABLE EXTRACTED FROM THE REGISTRATION SYSTEM.

Field	Description	Data type	Domain	Example
Name	Student name	string	A-Z	FAISAL ALI AHMED AJAJ
ID	Student's ID	Number	integer	20081234
Degree	Students program (major)	string	A-Z	Office management
S_Cr_hrs	Semester Credits Attended	number	0-21	9
C_Cr_hrs	Cumulative Credits Attended	number	0-180	30
S_GPA	Semester GPA	number	0-4	3.2
GPA	Cumulative GPA	number	0 - 4	3.5

TABLE 2
STUDENT TRANSCRIPT TABLE EXTRACTED FROM THE REGISTRATION SYSTEM.

Field	Description	Data type	Domain	Example
ID	Student's ID	Number	integer	20081234
code	course code	string	A{A*}d{d*}	CSA111
grade	grade obtained	string	A,B,C,D,F,B+, C+,D+,B-,C- ,A-,W,WF,IS,S,	A+
repeated	how many times this course was taken	number	0..3	0
Year	the date(year) when this course is registered	number	dddd	2008
semester	the semester when this course was registered	number	1..3	1

3.1 Text Processing Algorithm

The algorithm used to process the students' transcript file follows the same steps used in third generation languages compilers. It scans the file character by character, identifies tokens, and takes actions according to the recognized token.

Pseudo code

1- Open the transcript's file for input

2- while not end of transcript file do

3- read the file character by character and identify tokens

4- check if the token is one of the following identifiers: [course , GPA , name , status, degree, year, semester, passed]

5- for each identifier in Step 4, take the appropriate action to extract the required information.

Example of action
Theory:
let α , and β, represent any string, X,C and Y represent single non terminals (identifier), and β represents a terminal symbol (information[section, course, name, cumulative GPA]).

$\alpha \rightarrow \beta$
$\beta \rightarrow X \mid Y \mid C$
$X \rightarrow Name$
$Y \rightarrow ID$
$C \rightarrow GPA$

Application:
α→ Name: Ahmed Mohammed ID:1001234 GPA:3.2
Name → Ahmed Mohammed
ID → 1001234
GPA → 3.2

4. Advising and Graduation Processes

4.1 Data Description

The information derived from the registration system as described in Section 3 is mainly related to the student study progress as shown in Tables 1 and 2. The college offers 14 programs. In order to generate the list of courses required for students in different programs, the registration requirements for all programs must be input to the system. For each program, the list of courses along with their titles, codes, credit hours, and prerequisite details must be maintained as shown in Tables 3 and 4. Each student's study progress is compared with his program as explained in Section 4.2.

4.2. Process Description

The advising process starts by querying the existing registration system for student transcripts. For each student, a customized integrated web browser component gets the student's transcript as an HTML format which is then converted to text file format. The text file processing process is then invoked to convert textual data to database tables as shown in Tables 1 and 2.

TABLE 3
TABLE HOLDING DETAILS OF COURSES FOR A PROGRAM

Field	Description	Data type	Domain	Example
Code	Course code	string	A{A*}d{d*}	CSA111
Title	Course title	string	A-Z	Introduction to PC
Cr_hrs	Credit hours	number	0-6	3
Type	Type of the course major course, or service course	number	0-1	1
Lab_hrs	Lab hours	number	0-4	3

TABLE 4
TABLE OF COURSES AND THEIR PRE-REQUISITES

Field	Description	Data type	Domain	Example
Code	Course code	string	A{A*}d{d*}	CSA113
Pcode	Prerequisite code	string	A{A*}d{d*}	CSA112

The advising process is then invoked using database queries and data management procedures. Students' transcripts are compared against their programs and the list of courses for which a student must register in the new semester is generated. For a student, the process determines all unregistered courses and the prerequisite for each. For each unregistered course, if the prerequisite has been successfully completed the course is listed and can be registered. The process is then repeated for all previously unregistered courses and a list of courses in which the student should register in the following semester is generated. This process is repeated with every student. A sample report generated by this process is shown in Fig 1.

The student is considered to be graduated if he/she has successfully completed all courses in the program with a specific cumulative GPA and major GPA. The system checks whether the student meets the graduation requirements or not. Students who satisfy the study program graduation requirements are considered graduated and a special report for each student is generated as shown in Fig 2. The report shows all completed courses, the grade attained in each course, and the Cumulative and Major GPAs.

serial	ID	Name	Major	Course 1	Course 2	Course 3	Course 4	Course 5
1	19975003	FORAT M	Associate Diploma in Computer Networks	ENGLA 111	EEA 111	CEA 111	CSA 112	CEA 112
2	19975004	FATEMA	Associate Diploma in Multimedia Applications					
3	19975006	KHADIJA	Associate Diploma in Office Management	ARABA 111	MISA 233	MGTA 240	MISA 260	
4	19937498	ASMA S/	Associate Diploma in Office Management					
5	19899990	AQILAH E	Associate Diploma in Multimedia Applications	CSA 211	CSA 212	CSA 213	CSA 221	
6	19862482	KHADIJA	Associate Diploma in Office Management	MGTA 299				
7	19824974	ZAHRA A	Associate Diploma in Office Management	MGTA 299				
8	19787466	ZAHRA A	Associate Diploma in Chemical Engineering	EEDA109	MEDA161	CHEDA221	CHEDA242	
9	19749958	DUNYA H	Associate Diploma in Information Systems	MISA 299				
10	19712450	AMINA S	Associate Diploma in Information Systems	MISA 299				
11	19674942	ZAINAB I	Associate Diploma in Multimedia Applications	ENGLA 210	CSA 211	CSA 212	CSA 213	CSA 215
12	19637434	MARYAM	Associate Diploma in Web Design and Management	CSA 211	CSA 212	CEA 217	CSA 221	
13	19599926	AZHAR S	Associate Diploma in Multimedia Applications	CSA 299				
14	19562418	ZAHRA A	Associate Diploma in Civil Engineering	PHYCS 101	CEDA101	CEDA104		
15	19524910	JASIM M	Associate Diploma in Information Systems					
16	19487402	FATEMA	Associate Diploma in Office Management					
17	19449894	ZAINAB A	Associate Diploma in Information Systems	MISA 299				
18	19412386	RAJA KH	Associate Diploma in Information Systems	MISA 299				
19	19374878	ALAWI S	Associate Diploma in Multimedia Applications	CSA 113				

Fig. 1. Screen snap shot for courses to be registered next semester.

University Of Bahrain
College of Applied Studies
Associate Diploma in Office Management

Name :- ID 1989862

Course	Prerequesite	Exempted	Cr.	Grades	Repeated		Pnts	
Semester 1								
CSA 111			3	B-			2.67	
CSA 112			3	C+			2.33	
ENGLA 111			3	C+			2.33	
ENGLA 112			3	B-			2.67	
GSA 111			3	C+			2.33	
MATHA 111			3	B-			2.67	
semester 2								
ACCA 121			3	C			2	
ARABA 111			3	C			2	
ENGLA 120			3	B-			2.67	
MGTA 121			3	C			2	
MGTA 160			3	C			2	
semester 1								
MGTA 140			3	B-		F	2.67	
MISA 231			3	B			3	
MISA 233			3	C			2	
semester 2								
MGTA 222			3	B+			3.33	
MGTA 242			3	D+			1.33	
MGTA 240			3	D			1	
MGTA 237			3	C-			1.67	
semester 1								
MGTA 262			3	B			3	
MGTA 290			3	A-			3.67	
MISA 260			3	C-			1.67	
semester 2								
MGTA 247			3	B-		F	2.67	
semester 3								
MGTA 299			0	S			0	
	0	I	66	2	F	0	WF	0 E

C GPA =2.35 | M GPA =2.26

I have Checked This Graduation Application and Found it Satisfying The Graduation Requirements.
Completed Graduation Requirements

Advisor Name Signature .... Date

Fig. 2. Screen snap shot for graduates report.

5. Course And Sections Statistics Generation

5.1. *Data Description*

During the current semester, details of courses and sections to be offered for the next semester are requested by the registration deanship. In the current semester while students are attending courses, it is difficult to predict accurately which courses they will pass in order to determine the list of courses and their sections which must be offered for the next semester. Therefore, reading students' transcripts on its own as explained in Section 4.2 will provide us with incomplete information as details of courses in which students are currently registered must be taken into consideration. To read currently registered course details, the system accesses the students' current HTML formatted schedules as shown in Fig 3 using a customized integrated web browser component and converts it to text document format. The text document in return is parsed using the algorithm explained in Section 3.1 with a different set of identifiers such as ID, Code, Year, and Semester. This information is then stored in a separate table as shown in Table 5. In order to generate the statistics, it is considered that the students will successfully complete their currently registered courses.

Courses completed and currently registered are used as a base to generate courses to be offered and their section statistics. In order to specify courses for different programs, it is necessary to determine the program in which each student is registered. The data table shown in Table 6 holds the list of courses for which students should register along with the program in which the student is registered. The data table shown in Table 7 holds the list of courses for each program. Both tables are extracted as explained in section 5.2.

Current Schedule 2009/2

(20073454) MASAAD NAJI ALI MASAAD

Course	Crd	Sec.	Sat	Sun	Mon	Tus	Wed	Exam Information
CSA111	3	05	......	10:00-10:50 (B20-116)	11:00-13:30 (C20-106)	10:00-10:50 (B20-116)	......	21/06/2010 monday 11:30 - 13:30
ENGLA111	3	05	......	09:00-09:50 (B20-118)	....	09:00-09:50 (B20-118)	......	14/06/2010 monday 08:30 - 10:30
ENGLA112	3	02	......	11:00-11:50 (B20-120)	....	11:00-11:50 (B20-120)	......	16/06/2010 monday 01:30
GSA111	3	05	......	12:00-12:50 (B20-103)	....	12:00-12:50 (B20-103)	......	15/06/2010 monday 08:30 - 10:30

Total Credit Hours Registred = 12

Fig. 3. The current schedule web page snaps shot - registration system.

TABLE 5
TABLE HOLDING CURRENT SEMESTER SCHEDULE.

Field	Description	Data type	Domain	Example
ID	Student's ID	Number	integer	20081234
code	course code	string	A{A*}d{d*}	MGTA222
Year	the date(year) when this course is registered	number	dddd	2008
semes ter	the semester when this course was registered	number	1..3	2

TABLE 6
TABLE HOLDING COURSES TO BE REGISTERED FOR

Field	Description	Data type	Domain	Example
Stno	Student ID	number	dddddddd	20061234
Code	Course code	string	A{A*}d{d*}	CSA112
PID	Program ID	Number	1..20	5

5.2. Process Descriptions

The courses to be offered and sections generation process starts by storing the list of courses to be registered described in Section 4.2. A code for the program in which each student is enrolled is also stored in the table shown in Table 6. The list of sections for a course to be offered is then generated using the information stored in this table. Courses are grouped under a program and sections are grouped under a course. A limit for a section size is set e.g. 25. A report as shown in Fig 4 is generated showing the number of students in a program that can be registered in each course.

The main challenge in this process is students' grades for the current semester must be available. It is; therefore, better to execute this process at the end of the semester. To overcome this limitation and prepare the timetable for the coming semester before the end of the current semester, an assumption can be made that all students will pass all courses in which they are registered. Sections for the coming semester are then offered, bearing in mind that there will be a certain failure rate which could result in the possible closing of courses or sections. The process by which courses or sections can be closed must be simplified. Timetable adjustments can be performed once the current semester grades are input to the registration system and processing outlined in this section is repeated using fresh student information.

6. Conclusion

The paper presents an online advising system that is run by the department. The system is used to advise students enrolled in different programs on which courses they should register for in the next semester to stay on the correct study path and graduate within the expected time frame. A graduation report can be printed showing the list of courses a student has completed along with the list of courses to be completed if any. The system generates a list of courses and the number of sections for each course to be offered in the coming semester. Such information is based on the courses completed by the students enrolled in each program and on the remaining courses to be completed. The system eases advising and graduation processes, greatly improves accuracy, and helps in timetabling courses for the coming semester.

TABLE 7
TABLE HOLDING COURSES BELONGING TO A PROGRAM

Field	Description	Data type	Domain	Example
Code	Course code	String	A{A*}d{d*}	CSA112
PID	Program ID	Number	1..20	5

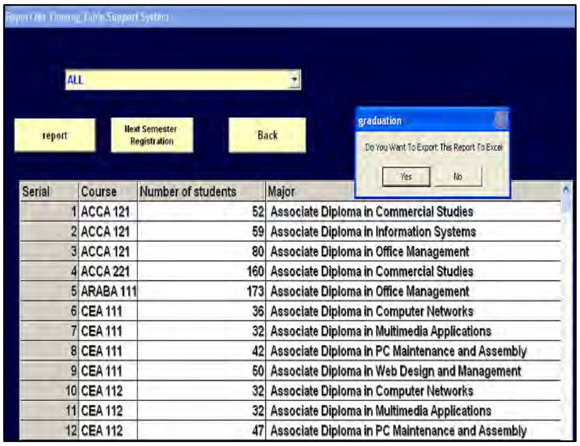


Fig. 4. Screen snap shoot for time table support system.

7. Future Work

The existing system is tailor made for the programs in our college. We plan to make improvements to the system so that it can be used easily by all other departments and colleges. We are also strongly motivated to produce a timetable and load allocation system that interacts with the system described here to generate a timetable that specifically suits students' course registration requirements and at the same time efficiently allocates courses to instructors.

Acknowledgment

Thanks goes to deanship scientific research at the University of Bahrain for supporting research published in this paper. We also would to thank Ms Alexandra Fraser for proof reading the paper and the anonymous reviewers for their comments and feedback. Finally, our thanks also go to the conference organizers for publishing our work.

References

[1] P. Bailes , X. Li, B. Lovell, S. Sadiq , X. Zhuo, "The Academic Advisor's Apprentice – An Automated Academic Advising System," CSEE Strategic Development Proposal.

[2] O. Marques, X. Ding, and S. Hsu, "Design and Development of A Web-based Academic Advising System," in *31st ASEE/IEEE Frontiers in Education Conference*, October 10-13, 2001 Reno, NV.

[3] R. M. Siegfried, A. M. Wittenstein, and T. Sharma, "An Automated Systems for Course Selection and Scheduling," *Journal of Computing Sciences in Colleges*, vol. 18, no. 3, pp. 17-25, February 2003.

[4] R. M. Siegfried, "FROSH: An Expert System for Freshman Advisement," *Journal of Computing in Small Colleges*, vol. 9, no. 2, pp. 88-98, November 1993.

[5] M. P. O'Mahony, and B. Smyth, "A Recommender System for On-line Course Enrolment: An Initial Study," in *RecSys'07*, Minneapolis, Minnesota, USA, October 19-20, 2007, published by ACM.

[6] D. Pokraja, and M. Rasamny, "Inetractive Virtual Expert System for Advising (InVEStA)," *36th ASEE/IEEE Frontiers in Education Conference*, San Diego, CA, October 28-31, 2006.

[7] V. R. Naini, R. S. Sadasivam, and M. M. Tanik, "A Web-based Interactive Student Advising System using Java Frameworks," *Proceedings of the SoutheasternCon*, pp. 172-177, *IEEE 2008*.

Using Computer Game to Share knowledge

[1]Mazeyanti M Ariffin, [2]Nazleeni Samiha Haron @ Baharon and [3]Mohd Izad Che Muda
[1,2,3]Computer & Information Sciences Dept
Universiti Teknologi PETRONAS
31750 Tronoh, Perak, MALAYSIA

***Abstract*— sharing knowledge is the key dimension of a learning organization. Organization that wants to sustain their competitiveness must encourage people to learn from each other or in other words to share knowledge among other. Variety mechanisms to share and transfer knowledge among knowledge workers are being used nowadays. These knowledge sharing mechanisms relatively depends on the nature and the type of knowledge to be transferred. Most organizations are using mentoring, face-to-face meeting, cross-functional and electronics means to share knowledge. Scholars claimed that simulation game can encourage human learning as players are able to make meaning during the play. This paper proposes to use simulation game as an alternative to share knowledge. A simulation game called FireHazard is developed and tested to selected participants. This paper also discussed on the development and the result obtained from the study.**

***Keywords*—simulation game, human learning, knowledge sharing mechanism**

I. INTRODUCTION

As inspired by [1], knowledge has become a new basic economic resource as compared to capital, natural resources or labor. Organizations start to realize the importance of knowledge [2] and its strategic impact [3,4,5] to sustain its competitiveness in a dynamic knowledge economy. As many organizations aim to increase profit by reducing their production cost, they made their operations globalize [6,7]. While this can reduce the operation cost, it imposes new challenge to organization. How to ensure knowledge that resides in the organization can be shared and transfer effectively despite their geographical location?

As such, researchers have proposed techniques and mechanism for effective knowledge transfer and sharing. One of the methods introduced is by using game and simulation. It has been proven that this method can aid in enhancing learning. Therefore, we have opted for this method as a mean for transfer and sharing knowledge on fire hazard. Fire hazard information is one of the crucial knowledge that must be possessed by all staff in any organization. This paper presents the theories behind knowledge sharing and computer game as well as the description of the game that has been developed to validate our study.

This paper is segmented into several sections : section II discuss on literature review to support the study, section III talks about methodology used, section IV discuss the result and discussion.

II. LITERATURE REVIEW

A. Taxonomy of knowledge

Referring to [8] knowledge is actionable information that leads to relevant action [9]. Reference [9] further added that knowledge is defined as human understanding of specialized field of interest that has been acquired through study and experience. An individual collect knowledge based on learning, studying, thinking and experience. People also get knowledge by familiarizing with situation, condition or problem in a specific domain or area. According to [9], knowledge includes perception, skills, training, common sense and experience. Knowledge helps people to make meaningful conclusion on certain problem or condition. It is know-how or familiarity of how to do something that enables a person to perform a specialized task [9]. Referring to [10] further added knowledge is the ability to turn data and information into effective action.

Knowledge can be categorized into two forms; explicit and tacit knowledge [11, 12]. Tacit knowledge is embed in human mind and stems to individual experience [9] , therefore it is hard to articulate or express. In contrast, explicit knowledge is knowledge that can be codified such as manuals, documents, reports, white papers and spreadsheets which are easier to share among individuals. ASHEN framework introduced by [13], splits tacit and explicit knowledge into six other types; documents, methods, skills, relationships, talent and experience.

B. Knowledge sharing

Knowledge sharing is defined as the transmission of knowledge from one source to another source and use of the transmitted knowledge [9]. He further adds that knowledge sharing only happen when there is transmission from the sender and absorption by the receiver. Referring to [14] support this claim by defining knowledge sharing as when the receiver's (e.g., group, department, or division) performance is affected by the experience of the sender. Reference [15] and reference [16] point out that knowledge sharing is a key of a learning organization. And learning process only occurs when knowledge in one part of an organization is transferred effectively to other parts and it is used to solve problems or simply to create new and creative insight. The success of knowledge sharing can be evaluated through changes in performance of receiver [14].

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_6,

C. *Knowledge sharing mechanisms*

Having known that knowledge can exist in many forms; organizations need to use different way or mechanism to share knowledge within the organization. It depends on the nature of knowledge. Reference [17] proposed that the type of knowledge sharing mechanisms is highly depends on the type of knowledge to be shared. As knowledge sharing mechanisms are categorized into 'reach', 'richness', 'tacitness' and 'embedness', mechanisms that has high level of 'reach' are more suitable to create awareness while mechanisms with high level of 'richness' are often more effective in transferring knowledge among individuals as shown in Table 1

TABLE 1: Determinants of 'Rich and 'Reachness' [17]

Mechanism' characteristics	Desired elements	General guidelines
Reachness	High number of receivers	How many receivers is the mechanism likely to reach at one time? What is the likely cost for potential receivers?
	Ability to overcome geographical barrier	How easy is the mechanism to reach receivers in different location? What is the cost involved?
	Ability to overcome temporal barrier	Is the knowledge stored in people or non-people medium?(people medium are assumed to be less capable in overcoming temporal barrier as they can be lost more easily when people leave the organization Is the knowledge with one or more people? (mechanism in which knowledge is spread to many individual is more likely to withstand turnover of people than mechanism which concentrate with only one person)
	Ability to overcome functional/departmental barrier	What is the likelihood of the 'outsider' coming in contact with mechanism?
Richness	Ability to transfer a lot of information at one time	How much information the mechanism can transfer at one time? What is the cost?
	Ability to transfer a lot of information of different nature at one time	What are the communication tools that likely to involve? Can the knowledge receivers meet face-to-face with the knowledge resource/originators?
	High interactivity between receiver and sender	Can the receiver and knowledge holder meet face-to-face? At what cost?

In some organizations, people share knowledge through transfer of people, forums (internal conferences/meetings), boundary spanners, periodicals, audits, benchmarking, best practice, guidelines and international teams [17]. Referring to [15] points out that mentoring, chat room, personal intranets and face-to-face conversations also being used by organizations to share knowledge. Thus, before organization invests on certain knowledge sharing mechanisms, they need to identify the goal and the nature of knowledge to be shared. The ultimate goal of this knowledge sharing mechanism is facilitate learning process, increase the absorption level of the receivers, and eliminate the geographical barrier. Many scholars like [18] claim that the effective tool to achieve this goal is to use computer-aided systems.

D. *Game and simulation*

Researchers such as [19] and [20] starts to realize the potential effects of games and simulation in instructional field years ago. According to [21], game consists of rules that describe allowable player moves, game constraints, privilege and penalties for illegal actions. By referring to [22] describe game as rule-governed, goal focused, microworld-driven activity incorporating principles of gaming and computer–assisted instruction while in contrast simulation is the one that consist real settings, participants' role, playing rule and scoring system [23]. By referring to [21], games and simulations carries distinct criteria and functionalities such as (1)observable characteristics used in games is more to 'drawing cards, moving pieces etc while simulations provide 'a scenario or set of data' which to be addressed by the players (2) the goal of game is to provide competition and winning state while in simulation participants are given serious responsibilities, and conducting feasible job procedures and possible consequences (3) event sequence in game is linear while simulation provides 'branching' or 'nonlinear' event (4)rules and regulations in games are unnecessarily to be realistic however those in simulations are authentic and closely related to real world.

There are lots of research being done in game and simulation as a method to enhance learning such as [24,19, 25,26, 27]. Reference [28] proposed that simulation can help people to learn without the need to develop costly training simulations. In other research conducted by [25], they proposed that computer simulation may be used to help in adults' learning skills particularly in business skills. Simulations and games are well accepted as a method that has impact on teaching and learning especially in military and aviation settings [27,28]. Reference [29,30,25], further support these claims by saying games and simulation are a powerful alternative to teach and learn as compared to conventional ways since this method is able to facilitate experiential learning. Other study done by [30] suggests that computer games help to enhance learning process and retention of knowledge in participants.

III. METHODOLOGY

Focusing on sharing knowledge activity and human learning, authors developed a simulation game called FireHazard using evolutionary prototyping phases. The software used was Game Maker, Game Maker Language (GML) Anime Studio, Adobe Photoshop, Audacity, PHP, MySQL and Apache HTTP Server.

A. Simulation game architecture

Figure 1 illustrates the architecture of FireHazard. It consists of four main components which are the game, knowledge repository, scoring system and score database.

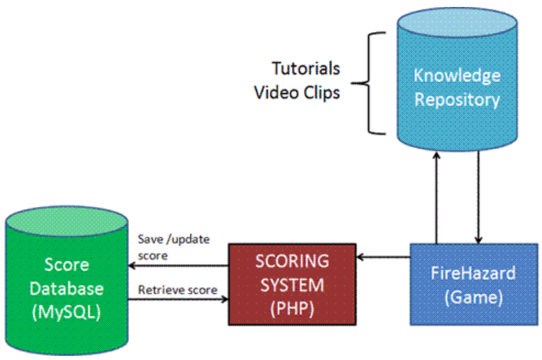


FIGURE 1: Architecture of FireHazard

Game is the main feature of the application that contains the client module. The client module provides the user interface of the application and is linked to knowledge repository and scoring system. Knowledge repository stores video and text tutorials regarding how to use fire extinguisher and how to put out fire properly. The stored information can be viewed via the Game interface. Furthermore, scores that are obtained from the game will be sent to scoring system for calculation using pre-determined formulas. The scores will then be stored in the score database. The overall score obtained by each player can also be retrieved and will be displayed at the Game interface. The summarized description of each component is discussed in Table 2.

TABLE 2: Description of FireHazard Main Components

Component	Description
Knowledge Repository	Consists of tutorials and video clips about how to put out fire using the correct techniques. The players have to read the tutorials and watch the videos before playing the game.
FireHazard game	The main portion of the FireHazard application. The game consists of three levels depicting three different scenarios.
Scoring System	The scoring system receives the scores from the game and calculates the total score using various pre-determined formulas.
Score Database	Stores the score of all players which consists of the subtotal score for each level and the total score for all three levels.

B. Game Description

FireHazard is a fire-fighting computer game with the aim to educate players on how to react in the case of fire. Players should be able to identify the type of fire extinguisher to be used to put off different type of fire. If players failed to use the correct fire extinguisher for the scenario, the fire will not fade away and deteriorate. The FireHazard game is purposely designed to simulate three different scenarios to maintain consistency. Scenario A depicts a situation of a Class A type of fire which caused by ordinary combustible materials such as rubbish and paper. Player must decide what to do in this situation and which fire extinguisher he must use to put off the fire based on the clue or hint given in the scenario. Scenario B of the FireHazard game depicts fire in a kitchen caused by Class A and Class C type of fire. Player should be able to identify and use fire extinguisher two different fire extinguishers to put off the fire. Player must think instantly as the fire is spreading fast and rapidly. Scenario C illustrates another different scenario which in this case in a chemical laboratory caused by Class D type of fire. If player use incorrect type of fire extinguisher, the fire will grow bigger and uncontrollable.

C. Play the game

Player is required to listen to set of tutorials and videos on how to use fire extinguisher and how to identify the type of fire. Later, players are required to play the FireHazard game to test their ability to apply knowledge that being transferred to them. Scores will be given if they manage to successfully apply what they have learnt and marks will be deducted if they failed in their attempt. Refer to Figure 2.

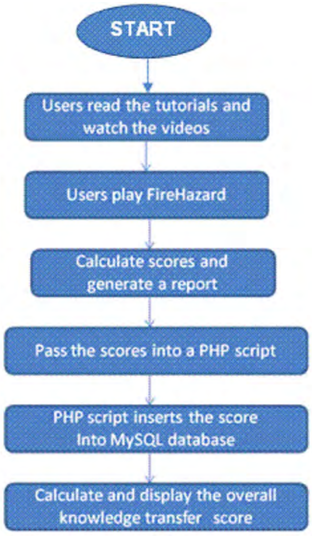


FIGURE 2: Game flow of FireHazard

D. Players

25 players were chosen for this study. These players were selected from Health and Safety department, lecturers and students who have no experience in playing the FireHazard game but have prior experience in playing other computer game.

IV. Result & Discussion

A. User Interface

FIGURE 3: FireHazard interface

FIGURE 4: FireHazard interface

Figure 3 and Figure 4 show sample interface of FireHazard simulation game. In this simulation game, players are being exposed to several types of fire; range from Class A to Class D. Three different scenarios are being developed to encourage players to apply their knowledge based on the previous encounter. FireHazard simulation game also provides hint and guidance to the players when they provide wrong answers.

B. Results

TABLE 3: result of playing FireHazard simulation game

Participant	Scenario 1			Sub Total	Scenario 2			Sub Total	Scenario 3			Sub Total	
	A1	B1	C1		A2	B2	C2		A3	B3	C3		TOTAL
1	40	40	20	100	40	40	20	100	40	40	20	100	100
2	40	30	20	90	40	30	20	90	30	30	20	80	86.7
3	30	30	10	70	30	30	20	80	30	30	20	80	76.7
4	40	40	10	90	40	30	20	90	40	30	20	90	90
5	40	40	20	100	40	30	20	90	40	40	10	90	93.3
6	40	40	20	100	40	40	20	100	40	40	10	90	96.7
7	30	30	10	70	30	30	10	70	40	30	10	80	73.3
8	30	40	20	90	30	40	20	90	40	40	20	100	93.3
9	30	30	20	80	30	30	20	80	30	30	20	80	80
10	40	40	20	100	30	40	20	90	40	40	10	90	93.3
11	40	40	20	100	40	30	20	90	40	40	10	90	93.3
12	30	30	10	70	30	20	20	80	30	20	10	70	70
13	30	30	20	80	30	30	20	80	40	30	20	90	83
14	30	30	10	70	30	30	20	80	30	30	20	80	76.7
15	40	40	20	100	40	30	20	90	40	40	10	90	93.3
16	30	30	20	80	40	30	20	90	40	40	20	100	90
17	30	30	10	70	30	30	10	70	40	30	10	80	73.3
18	40	40	20	100	40	30	20	90	40	40	10	90	93.3
19	30	30	10	70	30	30	10	70	30	30	20	80	73.3
20	30	30	20	80	30	30	20	80	40	30	20	90	83.3
21	30	30	20	80	30	30	20	80	30	30	20	80	80
22	40	40	20	100	30	40	20	90	40	40	10	90	93.3
23	30	30	20	80	40	30	20	90	40	40	10	90	86.7
24	30	40	20	90	30	30	20	80	40	30	20	90	86.7
25	30	40	10	80	30	30	10	70	40	30	10	80	76.7

Table 3 shows the overall scores obtained by all players for each scenario. The results exhibit that all 25 players obtain consistent scores throughout their games. Additionally, most of the players got more than 80 marks as the total score. Based on the results, we can infer that the players have successfully practice and applied the newly acquired knowledge obtained.

V. Conclusion

FireHazard is suitable to simulate 'real' situation in a controlled environment. As conducting the real situation is risky and costly, FireHazard simulation game able to provide 'artificial' experience to the players. By playing FireHazard it helps players to improve their learning process by allowing players to learn by taking responsibilities in performing specific job procedures and faced possible consequences in a controlled environment. Game and simulation provides experiential learning by giving opportunity to learn thorough experience and discovery.

Knowledge repository is chosen as the transfer method. This method of transfer is suitable for large audience because experts do not have to be available to perform the transfer since the experts in a large organization cannot afford to cater to all members in the organization.

References

[1] Peter F. Drucker, *Post Capitalist Society*, HarperBusiness, p 8, 1993

[2] Yozgat, U. Yönetim Bilişim Sistemleri. İstanbul: Beta Basım Yayım, 1998.

[3] Conner, K.R., Prahalad, C.K.. "A resource-based theory of the firm: knowledge versus opportunism". *Organization Science*. 7/5: 477-501. 1996

[4] Nahapiet, J., Groshal, S. "Social capital, intellectual capital, and the organizational advantage". *Academy of Management Review*. 40/2:242-266, 1998

[5] Flaherty, M. *Global Operations Management*, McGraw-Hill, Singapore 1996

[6] Yip, G. *Total Global Strategy*, Prentice Hall, New Jersey. 1992

[7] Tiwana, A. *The Knowledge Management Toolkit*, Prentice-Hall: Saddle River, 2000.

[8] Elias M. Awad, Hassan M. Ghaziri, *Knowledge Management* Upper Saddler River, New Jersey : Pearson Education, Inc., 2004

[9] Applehans, Globe, Laguero. *Managing Knowledge: A Practical Web-Based Approach.* 1999

[10] Polanyi, M. *The Tacit Dimension*, Routledge & Kegan Paul, London. 1966.

[11] Nonaka, I. and Takeuchi, H. *The Knowledge-Creating Company*, Oxford,University Press, New York. 1995

[12] Snowden D " *A Framework for creating a sustainable programme" CBI Guide to Knowledge Management*, London Caspian Publishing/Confederation of British Industry 1998 (Republish in Knowledge Management Year Book 1998 Butterworth Heinemann, April 1999

[13] Argote, L., Ingram, P., Levine, J. & Moreland, R. *"Knowledge transfer in organizations: learning from the experience of others,"* Organizational Behavior and Human Decision Processes, 82(1), 1-8. 2000

[14] Goh, S.C *"Toward a learning organization: the strategic building blocks"*, SAM –advanced Management Journal, Vol. 63 No. 2, pp. 15-22. 1998

[15] Garvin, D., *"Building learning organizations"*, Harvard Business Review, July-August, pp. 78-91.1993

[16] Chai, K H, Y S Shi and M J Gregory, *"Bridging islands of knowledge: A framework of knowledge sharing in international manufacturing networks In Managing Operations Networks"*, Proceedings of the EurOMA 1999 Annual Conference, 7-8 June 1999, Venice, Italy, edited by E Bartezzaghi et al., pp. 853-860, 1999

[17] Goh, S.C, *"Managing Effective Knowledge transfer:an integrative framework and some practice implications"*, Journal Knowledge Management, Vol 6 No 1. pp23-30. 2003

[18] P.S. Goodman, E.D. Darr, *"Computer-aided systems and communities: Mechanisms for organizational learning in distributed environments"*, Management Information Systems Quarterly, Vol. 22 pp.417 - 440. 1998

[19] Ruben, B.D, *Simulations, Game and Experience Learning,: the quest for a new paradigm for teaching and learning*, Simulation and Game 30(4) 498-505,1999

[20] Abt, C.*A Serious games*. New York: Viking, 1970.

[21] Gredler, M.E, *Educational games and simulations: A technology in search of a (research) paradigm*, In D Jonassen (Ed.) Handbook of Research for Educational Communications and Technology (pp521-540). New York: Macmillan 1996

[22] Driskell, J.E & Dwyer, D.J *Microcomputer videogame based training. Educational Technology*, pp11-16. 1984

[23] Christopher, E. M. *Simulation and games as subversive activities, Simulation and Gaming*, pp 441-455.1999.

[24] Schank, R. & Neaman, *A Motivation and failure in educational simulation design. In K. D. Forbus & P.J Feltovich (Eds), Smart machines in education: The coming revolution in educational technology*. Cambridge, MA, pp. 37-69, US: the MIT Press.2001

[25] Washburn, J., Gosen, J. *An exploration of game-derived in total enterprise simulations. Simulation & Gaming*, pp 281-296, 2001.

[26] Baker, E.L, & O' Neil, H. F. Jr. *Evaluation and research for technology: Not just playing around Evaluation and Program Planning*, pp 169-176. 2003

[27] O' Neil, H. F. Jr. , & Andrews, D. (Eds). *Aircrew training and assessment.* Mahwah, NJ: Lawrence Eribaum Associates. 2000

[28] Mayer, R. E., Mautone, P.m & Prothero, W. *Pictorial aids for learning by doing in a multimedia geology simulation game*. Journal of Educational Psychology, pp 171-185,2002.

[29] Rosenorn, T., & Kofoed, L. B. *Reflection in Learning Processes through simulation/Gaming. Simulation & Gaming*, pp432-440,1998

[30] Westbrook, J.I & Braithwaite, J. *The health care game: an evaluation of heuristic, web based simulation*, Journal of Interactive Learning Research , pp89-104,2001

Following the Paper Trail: Measuring the Economic and Environmental Impact of Digital Content Delivery

Vanessa K. Cox
Ruth C. May
Stanley L. Kroder
Geralyn McClure Franklin
University of Dallas
College of Business
1845 E. Northgate Drive
Irving, TX 75062

Abstract- We provide an overview of a framework for measuring the economic and environmental impact of digital content delivery currently being pilot tested in the College of Business at the University of Dallas. This framework is based on Elkington's Triple Bottom Line (TBL) model of sustainability which includes Economic (Profit), Environmental (Planet) and Social (People) dimensions [1]. We first describe the development of a rubric for quantifying online digital content followed by an overview of the process for assessing the environmental and economic effects of digital versus traditional paper delivery. Results are based on archival data generated with the rubric and survey data collected from 1,039 online students in the spring and summer semesters of 2009. We conclude with suggestions for establishing measurement models at other institutions where fostering partnerships between university administrators, faculty and students holds promise for increasing the sustainable impact of eLearning.

I. Introduction

While universities are increasingly focused on integrating sustainability concepts into their curricula, few have considered the potential impact that their eLearning programs may already be having on the environment and their bottom lines. Online classes result in some portion of course content *not* printed on paper, thereby reducing the carbon footprints of institutions and their students. Although academic institutions aspire to be recognized as champions of sustainability [2], recent progress made by corporations in measuring and reporting their sustainability efforts have not been matched by educational institutions, particularly in the domain of online course delivery. Consequently, the purpose of our paper is to present a framework for measuring the economic and environmental impact of digital content delivery currently being pilot tested at the University of Dallas (UD). Using archival and student survey data from the spring and summer semesters of 2009, we demonstrate the economic and environmental value of digital content delivery versus traditional paper and printing. The sustainable benefits of eLearning, at least in part, lie at the intersection of the Profit and Planet dimensions of Elkington's TBL as depicted in Fig. 1.

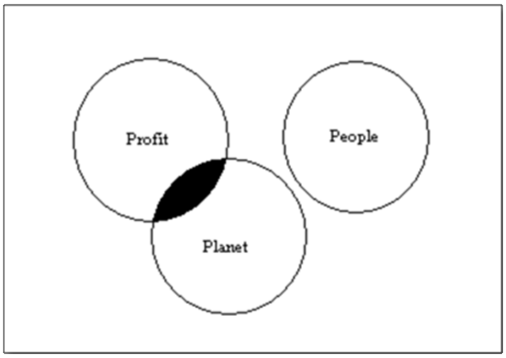


Fig. 1. Triple Bottom Line Dimensions:
University of Dallas Paper and Printing Pilot Study

II. Background

As concerns for the environment have increased among organizational stakeholders and competition for brand loyalty from customers has intensified, more firms are looking for ways to make sustainability a key pillar of their competitive strategy [3]. The definition of sustainability has evolved over the past 20 years, but is most often linked to the Brundtland Report issued in 1987 which proposes that sustainable development "meets the needs of the present without compromising the ability of future generations to meet their own needs." [4]. While institutions of higher learning have generally lagged the private sector in using sustainability for competitive advantage, recent progress by colleges and universities to close this gap between industry and education demonstrates a growing awareness among higher education administrators of the potential benefits to be gained by making sustainable policies and practices a centerpiece of their strategy. Even AACSB International – The Association to Advance Collegiate Schools of Business, the premier accrediting body for business schools globally, has been promoting sustainability and business school leadership in this effort with speakers and sessions at various conferences and seminars as well as its own Sustainability Conference [5]. Yet what cannot be measured

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_7,

cannot be managed effectively, much less used for strategic advantage.

The Graduate School of Management at UD, now part of the College of Business, has been involved in distance learning since 1969 beginning with a regional microwave video network. As the Internet emerged in the 1990s, the school leveraged the new technology by launching three online credit-bearing courses. This initial offering has evolved over the past decade to include the full Masters of Business Administration (MBA) core curriculum and a dozen specialized MBA concentrations, with no residency requirement. The expansion of the college's online delivery has provided an opportunity to reduce costs associated with traditional paper and printing content production and to reduce the carbon footprint of the institution. In early 2009, a pilot study was initiated to document these savings which are a natural byproduct of online education. The following section describes the process used to account for digital content **not** printed and the resulting savings in dollars and carbon emissions.

III. METHODOLOGY

A. Sample

The sample consists of online MBA students in the spring and summer 2009 semesters. While traditional brick and mortar MBA classes in the COB have companion websites for each class section, neither the digital content nor the students in these on-ground classes were included in this study. Rather, only fully online MBA students were invited to participate in an automated survey as part of a broader project focused on the impact of eLearning on all dimensions of the TBL model, much of which extends beyond the scope of this paper.

A week prior to launching the survey, students were invited to participate in the study via a notification message in the online course system. The online survey was inserted into the opening frame of one online course per student regardless of the number of online courses they were taking to avoid duplicate responses. Participation in the survey was optional. Out of 644 online students invited to respond in the spring 2009 semester, 539 provided complete data resulting in a response rate of 84%. Of the 621 online students enrolled in the summer 2009 term, 500 provided useable responses for a response rate of 81%. The combined sample from both semesters totaled 1,039 for an overall response rate of 82%. Across both semesters the student sample ranged in age from 20 to 63 with an average age of 37. Fifty-one percent of the respondents considered the Dallas /Fort Worth (DFW) metroplex to be their home base and 49% percent of the respondents resided outside the DFW metropolitan area.

B. Data Collection

In addition to the student survey data, archival data was generated using an original rubric designed to count the number of virtual pages available to students in all COB online courses. In spring 2009, 48 online sections were offered, and in summer term 44 online sections were conducted. Thus, the archival base used to compute total digital content for the COB consisted of 92 online course sections for spring and summer 2009 combined.

1) Digital Content: Calculating the total number of virtual pages required establishing a definition of what constitutes a virtual page as well as specific guidelines to ensure a consistent, accurate page count. We defined a virtual page as either a) actual pages in a paginated document or b) the equivalent number of pages for scrolling digital content in a Microsoft Word document. The following items were considered when counting the virtual pages in each online course.

- *Definitions:* The total number of virtual pages includes **obvious** printable downloads, **nonobvious** printable digital content and **nonprintable** digital content. Examples of obvious printable downloads include syllabi and XanEdu (vendor) course packs. Nonobvious printable digital content includes items such as chat sessions as well as directives to students in digital format. Nonprintable digital content includes grade book and exam items protected by plug-ins that prohibit the printing of designated content.

- *Guidelines:* Specific counting protocols of course content items were established and applied by a single counter in both semesters to ensure consistency. This approach reduced the risk of double counting items such as graphics, pictures or links. eBooks or external content located on other web sites were excluded from the page count.

- *Multiple User Access:* Once the total number of virtual pages per course was determined, we multiplied this figure by the number of students enrolled per course (i.e. total course seats) to determine total digital content provided in each online course. This approach is consistent with considering the sum of all copies of a traditional course textbook sold to every student purchaser versus assuming that a single copy of a textbook is shared by all students in a class. This perspective represents more accurately the total amount of digital content provided by the institution in a single semester.

2) Printing Behaviors: In order to determine the number of virtual pages offered by the COB that ultimately remained virtual (i.e. never got printed), students were asked about their printing behaviors in their online courses as part of the student survey. Specifically, students were asked to approximate the percentage of obvious downloadable documents and nonobvious printable content that they would typically print in an online class. In both document categories they were asked to select a percentage that they typically printed across all online courses from a drop down list ranging from zero to 100 % in 10% intervals. We elected to ask students about their printing behaviors in general rather than ask for a specific page count or rate for a specific class because we believed that the probability of getting reliable responses about printing habits at such a level of specificity was relatively low and would be more likely to reduce our response rate to the survey.

Students were also asked "What type of printer do you typically use for printing course related materials?". They could select one of the following: small desktop printer, shared workgroup within their company or work group, or a business center printer such as Fed-Ex Kinkos or a hotel.

IV. Results

A. Digital Content

Table I summarizes the virtual page count in all categories: obvious printable, nonobvious printable, nonprintable and total virtual pages. For both terms combined, the college provided over 1.5 million pages of digital content. However, the net number of virtual pages that remained in digital format (i.e. were not printed by students) could only be derived by applying the data captured on student printing behaviors. Specifically, students in the spring 2009 term printed an average of 53% of obvious downloadable pages and 30% of nonobvious printable content. Students in the summer 2009 term printed an average of 49% and 32% in the same categories, respectively.

The net results for the institution are illustrated in Fig. 2 where 543,275 pages remained virtual in the spring semester and 320,028 remained virtual in the summer term. Thus, a total of 863,303 pages of digital content provided by the institution remained in digital format, with their corresponding economic and environmental consequences, even after students' printing behaviors were taken into account for both semesters.

TABLE I
Rubric Summary
For All Online Courses

University of Dallas Term	Obvious Printable Pages	Nonobvious Printable Pages	Nonprintable Pages	Total Virtual Pages
Spring 2009	544,151	399,593	7,808	951,552
Summer 2009	421,757	147,722	5,381	574,860
Year to Date	965,908	547,315	13,189	1,526,412

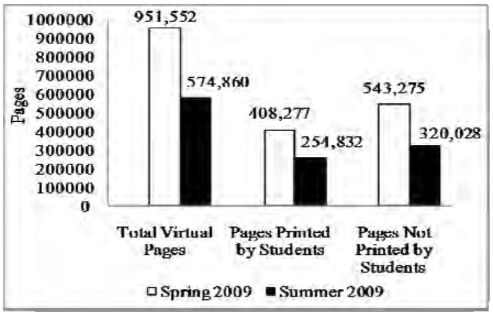


Fig. 2. Printing Behaviors

B. Economic and Environmental Consequences

Providing digital content versus traditional paper and printing delivery results in trees not destroyed to produce paper, energy not used to print content on paper, and carbon dioxide not emitted in the printing process. Collectively, these outcomes result in economic savings and environmental benefits for the planet. To quantify these benefits, we relied on the Hewlett-Packard (HP) carbon footprint calculator.

HP has a strong focus on global citizenship and sustainability with corporate goals that pertain specifically to environmental outcomes [6] [7]. Its printer division is a leader in the imaging market and has a "green" focus in its product and delivery. The HP Carbon Footprint Calculator is designed to assess printer energy consumption and associated carbon output as well as quantity of paper used and related monetary costs. The calculator generates estimates of energy consumption during use of a printer and emissions of carbon dioxide from production of estimated volumes of paper consumed during printing [8]. The savings derived by the calculator provided explicit figures for the study with the exception of the CO_2 emissions which had to be converted from pounds of CO_2 to a Metric Tonne Carbon Equivalent (MTCE), the standard unit of CO_2 emissions that we adopted for this study [9].

Table II provides the gross institutional savings in kilowatts, tonnes of CO_2 emissions, pounds of paper and dollars generated across all online courses in spring and summer 2009. While the savings depicted in Table II are real savings from an institutional viewpoint, the printing habits of students must be taken into account in order to determine the net real environmental benefits of the digital content delivery. After accounting for both printing habits and types of printers used by students, the net results for the planet remain positive as illustrated in Table III.

TABLE II
Carbon Footprint Calculator Results
For Total Digital Content

Gross Institutional Savings			
Savings In:	Term		Year to Date
	Spring 2009	Summer 2009	
Energy(kWh)	154,991	143,654	298,645
Carbon Emissions- CO_2 (tonnes)[a]	103.16	93.31	196.47
Paper (lb)	10,462	6,327	16,789
Energy and Paper (USD)	$ 20,243	$17,098	$ 37,341

TABLE III
CARBON FOOTPRINT CALCULATOR RESULTS AFTER ADJUSTING FOR STUDENT PRINTING

NET ENVIRONMENTAL BENEFITS			
Savings In:	Term		Year to Date
	Spring 2009	Summer 2009	
Energy (kWh)	33,538	31,949	65,487
Carbon Emissions-CO_2 (tonnes)[a]	24.85	22.20	47.05
Paper (lb)	5,978	3,531	9,509

[a] A tonne (t) or Metric Tonne Carbon Equivalent (MTCE) is a measurement of mass equal to 1,000 kg or 2,204.622262 lbs, or approximately the mass of one cubic metre of water [10].

In summary, for the first eight months of 2009, after accounting for student printing behaviors, the digital delivery of all online content in the COB resulted in 65,487 kilowatts of energy not used, 47.05 tonnes of CO_2 not emitted, and 9,509 pounds of paper not printed for a total institutional savings of $37,341.

V. CONCLUSIONS

This paper reported on the results to date of a pilot study currently underway in the College of Business at the University of Dallas to develop a framework for measuring the economic and environmental impact of digital content delivery. While many universities have stepped up efforts to integrate sustainability concepts into their curriculum and their on-ground campus operations, few [11], if any, have attempted to quantify the positive environmental effects that their eLearning operations are generating while producing cost savings as well.

Using survey data from 1,039 online MBA students in the spring and summer semesters of 2009 and archival data created via virtual page counts, we found positive economic and environmental benefits from digital delivery for the first eight months of the year. Cumulatively, digital delivery of content resulted in 298,645 kilowatts not used, 196.47 tonnes of CO_2 not emitted and 16,789 pounds of paper not printed for a total of $37,341 saved by the University of Dallas.

After taking into account the printing behaviors of online students, which is a more conservative measurement approach, the online operations of the COB still produced environmental benefits totaling 65,487 kilowatts saved, 47.05 tonnes of CO_2 not emitted and 9,509 pounds of paper not printed. The savings in carbon emissions alone equates to the following:

- Approximately 10 Round Trip flights from New York to Los Angeles **not** taken [12].
- Nine average passenger vehicles **not** driven for one year [13].
- 2.35 football fields of trees in a section of the Amazon Rain Forrest **not** cut and burned [14].

While the measurement process described here makes an incremental contribution to the challenge of measuring and reporting the sustainable impact of online delivery, it is subject to the following limitations:

- The measurement model does not take into consideration all aspects of digital content delivery at UD – COB (i.e. Hybrid or blended learning as well as web enhanced instructional companion elements) so the results are understated.
- The pilot study is located at a private university in a large metropolitan area, and thus the results are institution specific.
- The digital count of virtual pages included only fully online courses in a Masters of Business Administration program delivered in an asynchronous delivery model, and therefore, may not translate well to undergraduate programs with more synchronous content and interaction.
- While the survey response rate was excellent from graduate level students at UD, the same level of interest or participation may not be replicated with an undergraduate population.

Despite these inherent limitations, the pilot program demonstrates the value of considering economic and environmental consequences of digital content delivery for the institution and the planet. Moreover, while the pilot framework described here is only one approach to measuring the sustainable impact of online learning, our intention is to provide an initial blueprint for other educational institutions to follow in:

- Determining the volume of digital content they provide and the corresponding economic benefit.
- Determining the offsetting effects of student printing behaviors on the institution's digital content delivery strategy.
- Determining the combined effect of institutional and student choices on the broader environment.

University administrators, faculty and students have the opportunity to positively influence the economic and environmental dimensions of the TBL with regard to digital content. Specifically, administrators can be more proactive in adopting institutional strategies to increase online offerings in various academic units and can mandate the continuous tracking of benefits and costs associated with digital delivery. Administrators can also consider establishing green "stretch goals" for their online operations on both a gross and per capita basis.

Students have a significant role in determining how much of an institution's digital content remains virtual as a result of their printing behaviors. Therefore, it is important to inform students of the power that they have to reduce carbon emissions simply by becoming more aware of how much they print and how their printing decisions have direct environmental consequences. A leading edge example of institutional influence on student printing behaviors is a pilot program involving six colleges and Amazon.com. The partnership provides designated students at the pilot universities with a Kindle DX and access to required course readings. The e-book reader is designed to meet the needs of students by offering a larger screen and searchable content embedded within required reading materials while motivating them to reduce their printing of reading materials [15]. In general, today's students are likely to be more environmentally conscious than their predecessors [16], but they cannot contribute to a solution if they are not made fully aware of the problem and their potential to influence its outcome.

Likewise, instructors should be encouraged to focus on their roles in the sustainable efforts of their institutions by rethinking their design of course content in keeping with the "less is more" proposition. For instance, instructors should consider the following: How effective is pure text content and how is it likely to be perceived by students, particularly those in the Gen Y demographic? Recent studies show that Gen Y individuals prefer learning through multi-media and multi-tasking designs rather than traditional text delivery [17]. With this reality in mind, instructors should consider what might be more effective pedagogical options relative to text for increasing student engagement and reducing the temptation to print. These options might include relevant videos and student-directed content "scavenger hunts" based around a central theme or problem which prompts students to focus on collecting and analyzing virtual content as opposed to printing and memorizing basic text. This approach to content design is consistent with John Dewey's philosophy which is, "*Give the pupils something to do, not something to learn; and if the doing is such a nature as to demand thinking; learning will naturally result [18].*"

In conclusion, our results demonstrate the value of not just following the paper trail, but in documenting and measuring it in order to better guide organizations in achieving sustainable strategies. We hope that these initial efforts will prompt greater consideration and discussion of the potential benefits of digital content delivery and the corresponding challenges involved in measuring them.

ACKNOWLEDGMENTS

The authors of the paper thank John Frey, Americas Sustainability Executive with Hewlett-Packard Company. Under Frey's leadership, Hewlett-Packard has been named one of the 100 most sustainable corporations in the world by the Global 100 each year since the inception of the list. In addition, HP continues to be listed on key sustainability indices including the Dow Jones World Sustainability Index and the FTSE4Good. Mr. Frey provided important input and clarification related to the methodology of the Hewlett-Packard carbon calculator and the paper and printing aspects of the study.

The authors also thank Jeff Borden, Senior Director of Teaching and Learning for Pearson-eCollege. Mr. Borden has led his company's global efforts to deliver high-quality online courses for hundreds of higher educational institutions since 2002 and he provided valuable advice to our team during the design of the survey and data collection process.

Finally, the authors acknowledge with grateful appreciation the contributions of Nestor Via, Graduate Assistant in the online learning business unit of the University of Dallas, College of Business. Mr. Via conducted the count of all virtual pages in both spring and summer terms to ensure the integrity and consistency of the data.

REFERENCES

[1] J. Elkington, "Towards the sustainable corporation: Win-win-win business strategies for sustainable development," *California Management Review*, 36, no.2, pp.90-100, 1994.

[2] D. Ferrer-Balas, J. Adachi, S. Banas, C.I. Davidson, A. Hoshikoshi, A.Mishara, Y. Motodoa, M. Onga, M. Ostwald, "An international comparative analysis of sustainability transformation across seven universities," *International Journal of Sustainability in Higher Education*, vol. 9, no.3, pp.312, 2009.

[3] M. Porter, M. Kramer, "Strategy & Society: The link between competitive advantage and corporate social responsibility," *Harvard Business Review*, pp 1-2, December, 2006.

[4] W. Adams, "The future of sustainability: rethinking environment and development in the twenty-first century," *The world conservation union*,, online, last updated May 2006, http://www.iucn.org.

[5] B-school facilities resource center, online, http://www.aacsb.edu/resource_centers/fcilities/buildinggreen.asp .

[6]HP Global Citizenship Report: Environmental sustainability, online, September 24, 2009, http://www.hp.com/hpinfo/globalcitizenship/gcreport/envsustainability.html .

[7]Letter from Mark Hurd, HP Chairman, CEO and President, http://www.hp.com/hpinfo/globalcitizenship/gcreport/intro/ceoletter.html

[8] HP Carbon footprint calculator for Printing, http://www.hp.com/go/carbonfootprint .

[9] MtCO2- Metric Tonne Carbon Equivalent, online, 10/06/2009, http://www.mtco2e.com .

[10] Tonne, online, http://en.wikipedia.org/wiki/tonne .

[11] R. Roy, S. Potter, "Designing low carbon higher education systems: Environmental impacts of campus and distance systems," *International Journal of Sustainability in Higher education*, vol.9, No. 2, pp. 117-118, 2008.

[12] C.Dunn, "Reduce your carbon footprint by half in three steps: Plant Green", online, July, 2009, http://planetgreen.discovery.com .

[13] Emission Facts: Greenhouse Gas Emissions from a Typical Passenger Vehicle- US EPA, online, July, 2009, http://www.epa.gov/OMS/climate/420f05004.htm .

[14] How your pollution is calculated – Fight Global Warming- Environmental Defense Fund, online, July, 2009, http://fightglobalwarming.com .

[15] S. Peters, "Universities turn to Kindle- Sometimes to save paper," *New York Times*, online, July 30, 2009, http://greeninc.blogs.nytime.com/2009/07/30/universityies-turn-to-kindle-sometimes-to-save-paper/ .

[16] J. Pryor, S. Hurtado, J. Sharkness, W. Korn, L. DeAngelo, L. Romero, S.Tran, "Political engagement among college freshmen hits 40-year high," *Higher Education Research Institute (HERI): The American Freshman: National Norms Fall 2008*, online, http://www.gsei.ucla.edu/heri/pr-display.php?prQry=28.

[17] C. Dede, "Planning for neomillennial learning styles: Shifts in students' learning style will prompt a shift to active construction of knowledge through mediated immersion," *Educause Quarterly*, pp. 8-9. 2005.

[18]S.Khurana,About.com,online, http://quotations.about.com/od/stillmorefamouspeople/a/JohnDewey1.htm .

Context Driven Personalized e-Learning Environment Provided as a State Service

Deniss Kumlander
Department of Informatics, TTU
Raja 15, 12618
Tallinn, Estonia
e-mail: kumlander@gmail.com

***Abstract*—The main topic of this paper is the life-long education. The rapidly evolving world permanently challenge both individuals and governments as the productivity and educational level of the state work-force is one of the most important factors of the long term success in both financial and social sectors. The paper describes an e-education system that should be funded by the state and exposed to all citizens in order to constantly upgrade their skills. The broad potential auditorium and targets of such system define certain unique requirements that should be met by the context driven personalized approach of delivering information as proposed in the paper. It makes the process of acquiring knowledge simpler for all categories of users and stabilizes the virtual learning environment from the data protection point of view.**

Keywords-e-education;artificial intelligence; life-long learning; state services; context.

I. Introduction

There are two different approaches to the educational system in the modern society. One of those is university centric suggesting to concentrate only on educating young people. It proposes that correctly organized educational process gives enough skills to upgrade knowledge later on without mentors. Here the state builds up centers of massive education for young people and fund those also as research facilities in order to concentrate science in one place and produce this way highly skilled work-force, which is so much required by the modern world [1]. This approach has both positive and negative sides [2, 3]. For example it was criticized as been mostly looking into the past accumulating the knowledge in order to spread it without major updates of courses during sufficient periods of time [3]. Universities do argue that this is much more stable method of educating as proper understanding of fundamental science (basics, techniques) will give to students a good base for acquiring whatever additional knowledge in the future [5]. Having agreed with it, we still like to highlight the sufficient gap between the knowledge obtained in universities and the knowledge students will need dealing with specific tasks at their future workplaces. Moreover the ability to acquire new knowledge in the future doesn't necessary mean that the person will do that. It could be a problematic or extremely time consuming for him without having a good understanding of what is going on in the modern technology and having to parse a huge amount of information sources. Notice that the last mentioned tasks are tasks of the teacher in normal circumstances, who should concentrate the knowledge before delivering it to students. Therefore the university approach doesn't work very well in many cases of the modern, rapidly changing world. Constant technology, science, best practices, standards and so forth changes cause various types of unemployment, low productivity of work-force since skills and knowledge become outdated very quickly. This is a serious challenge for all involved parties: society, business and individuals and therefore it should be solved together in a synergy.

The concept of the life-long education is the first step on this road. Unfortunately in many cases it is left to individuals to follow it, while others do minor investments into that mostly motivating people to continue the education on constant basis. In the result many people spent their time trying to find and acquire knowledge in parallel instead of following mentor advices consuming it in a prepared form. Fortunately everything started to change (so far slowly) and the importance of an electronic learning as an alternate solution to the traditional education has been recognized. The quick development of electronic channels and world globalization [6, 7] help us to organize now the re-education much better than it sued to be in the past. Unfortunately those attempts mostly local and lack the strategic, global power that could be provide by a state financing and driving the development of such virtual education environment. At the same time, upgrading to the new level of virtual education means also facing completely new challenges in process of building such systems. There are already millions of students attending e-courses in developed countries and this number is growing permanently. They come from very different education levels, age and society groups. Therefore the online education based on average student doesn't work as well as it was hoped and requires personalization of the delivered content. Fortunately computers and artificial intelligence give us just enough tools to deliver such personalized content as described in this paper.

II. Personalized e-Education provided as a State Service

A. Auditorium

The kind of the system aimed in the paper has certain difference from other e-Education systems described a lot in the literature and the most important one is the targeted auditorium. If the standard literature case is moving the educational process from offline to online aiming still the

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_8,

university courses and students, then the state service is mainly directed to broader auditorium of citizens of different ages. So, here e-learning participants can be those who completed the last study long time ago, people who spent a lot of time on other primary activities, individuals who wish to obtain modern knowledge or acquire a completely new profession. Moreover we should still count with potential participants from universities or even schools. Therefore it is important to personalize the delivered courses as much as possible it order to increase a probability of successful acquisition of knowledge by each participants.

As a concluding note for the auditorium description we would like to mention that all this defines a challenging task for the system to acquire the context of students in order to personalize delivered information correctly.

B. Targets

The high level task of the proposed virtual educational system is to serve citizens of a state looking for possibilities to upgrade their knowledge. Besides, the system should be intelligent enough to find a knowledge delivery workflow suiting for a particular individual basing on her previous experience, background and availability.

All this must be explained in more details in order to identify correctly the scope and complexity of the goal. First of all such a virtual environment should be a proactive way of a state to deal with a potential unemployment and prevent the drop of workforce skills. The reactive approach is usually too expensive due a need to pay social benefits and have a low level of technological development loosing competition to other states. All this expenses are carried by the entire society in form of different taxes and the government task is to avoid them if possible. That is why the paper proposes the system to be developed as a service provided by the state. It generally means that it should be

1. Supported by the state (financing);
2. Advertised by the state;
3. Integrated into the general state infrastructure and made available basing on dedicated infrastructure elements implemented by the stated.

If the first two elements are easy to understand then the last one can easily be missed. Unfortunately it is not possible to build such a system from scratch. It should be an evolutional approach when the educational system follows other state services and relies on the same delivery infrastructure in order to make it available (in both meanings: ability to access and a constant availability). The educational content (data flow) is much bigger than other information service like registers and so demands sometimes completely new infrastructure system to be built up. Fortunately state usual can afford financing and supporting that considering the social cost it will carry vice versa.

Moreover the system should not unify or demand too much from consumers in terms of spent time and required knowledge level in order to participate. An ideal life-long education orients on gradual upgrade of work-force skills organized in parallel with the main participants' activities at work or home. Therefore we should foresee different level of education and previous experience of users. Besides the ability of individuals to acquire knowledge is known to be sufficiently varying dependently on their age and it makes students also different. Finally it should be possible to define any course on packets suitable to students' availability as highlighted earlier and be available online on demand.

C. Responsibilities

It is incorrect to assume that the organizer of the virtual learning environment is responsible only for the content of courses, i.e. quality of the learning process. Unfortunately there are a lot of other responsibilities, which should be taken into account building such a service.

First of all the service operates with information about the user and in many cases it is more just a username and a password. Notice that here we don't mean only the protection of the data in the communication channel between the e-learning server and client computers. The personalization of the system means that it should collect a lot of context about the consumer and all this information is stored on the server in forms of different logs, patterns and so forth. Moreover even the information on which courses were taken, when and what marks she got should be classified. Only the person (i.e. owner) could expose this information, for example temporary to the employer, and nobody else.

Secondly the course authors are usually also highly interested in protecting the content of lessons as much as possible. So, there is a clear demand to build virtual environments that will be both easy to consume by students and hard to misuse or replicate without the author permission. It is quite a challenge task and the easiest way to protect the content is to set a copyright on this. Unfortunately it is not enough in many cases, so authors will be thankful if the system will contain other protection mechanisms.

The protection of such private data is the most important challenge for the system developers. They need to ensure that data neither can be stolen nor made available outside due any kind of errors or backdoors of the system.

III. General Model

This chapter continues the description of the virtual learning environment introduced on our last paper [10] and proposes additional artificial intelligence elements supporting the personalization based on the acquired context. Therefore in the beginning an overview of the system is given to describe the topic defined so far: prerequisites for building such a system and interfaces of then model.

A. Prerequisites

The prerequisites will be described basing on a particular example of one European country – Estonia. The country has an informational system available to citizens to request their information hold by different state registers. The system is developed centrally by a special agency and mostly concentrates information available internally from different departments like Real-estate agency, Police, Automobiles registry and so forth. Moreover the system does allow citizens to send requests and documents to the state in order to obtain a new passport for example. All departments of the state should process such electronic documents signed by so called “ID-card” and there should be no difference between paper and electronic channel of communication. The information system has a single entry point.

This is an example of the system that normally should be in place before building the state-wide fully functional educational system. Here we would like to list some reasons why we believe such system is required:

1. The process of delivering information to citizens is verified and in use, so citizens know how to find and access it;

2. Citizens do trust to such system much more than to any other especially if the state is taking seriously the protection of private data;

3. The security (log in) is well established and is normally supported by some kind sophisticated mechanism of passwords distribution (in Estonian case: ID-cards issued by the state, which are equivalent to passports, i.e. well protected including the law);

4. The developers’ team is acquired and educated; relationship to external contractors (including the internet providers) is well established.

B. *Overview of model interfaces*

Such a broad system like described here has several complex interfaces designed to different type of actors.

1. Common part.

   First of all each interface does contain the security elements that can be inherited from the pre-system like the one described earlier. Secondly, there are a set of databases on the back end that contain course materials, results and other personal information on participants and so forth. Thirdly there are infrastructure elements that do support the communication like a load-balancing server, which should spread the load between a set of servers, since the state wide educational system is likely to be heavily used. Finally there should be a communication server that should allow communication between all involved parties outside the learning process in order to provide a) general feedback on courses, b) define interests from business community to design courses or stimulating participations – for example a job advertisement defining passing certain courses as a minimum requirement and so forth.

2. Courses’ authors interface.

   This interface allows building and uploading course materials into the system following certain system standards. If the author is not interested in personalization then those standards are quite simple and define types of files, structuring and examination rules. At the same time the supervisor of this system should stimulate producing personalized courses trying to keep simple courses set on minimum level and make this route available only for distinguished lectors. The interface should contain also templates making the course building process both simpler and standardized.

   Notice that for the simplification we assume that all courses do not require a mentor to help students to acquire the knowledge, although this element can also be included if required.

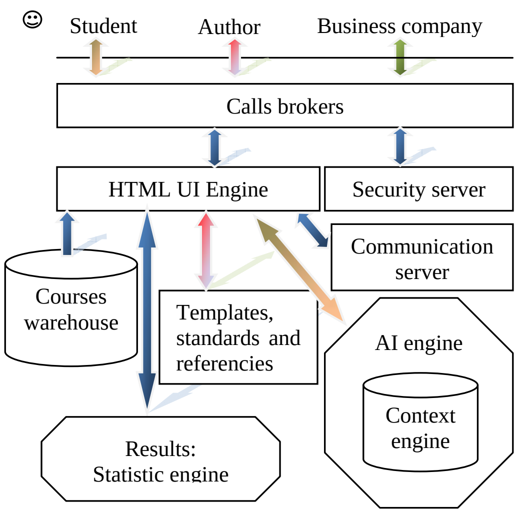


Figure 1. General model

3. Consumers (students) interface

   The nature of this interface is obvious: students do participate in courses by passing different topics of step by step and (optionally) pass an exam in the end. This is the interface we are going to personalized first of all by using AI and context engines shown on the figure 1.

4. Third-party interface

   The system should not be restricted to the course (teacher) – student communication. There are other groups of people including business community that may be interested in this process retrieving information if it is exposed or providing requests. This is a way job seekers could expose their professional level to companies. It is also an interface for companies to stimulate creation of a work-force they do require by motivating authors to create certain courses and students to participate and obtain high marks.

IV. Context driven personalization

A. Personalisation using context

The context-aware system is a very popular topic nowadays in different artificial intelligence systems starting from web pages up-to automobile user interfaces for drives.

Generally, context-aware system development means that the computer reaction can be organized using more than just the direct data input. Instead computer system should consider circumstances under which the input is provided sensing the environment in which the request is formulated and form the response basing on those parameters [10]. Some authors say that it is nothing else that a system ability to adapt to specific circumstances under which the system operates or reacts on the request and ability to collect and recognize the context information [11, 12]. Therefore the context-aware approach is mainly used to increase relevancy of the response (service) provided by the system and is targeted to bridge the gap between formal and human to human type interaction [13].

Unfortunately majority of scientists do restrict themselves with the most obvious context they can acquire – the location. Fortunately for us it is not the only one and the virtual education system is the perfect example of a system that could obtain and process other types of context in order to make the process of education very personal.

The previous experience and educational level can be identified as one context that can be used to build up personal courses. It can be acquired in a form of tests and questionnaires where a student defines both her current knowledge and topics she wishes to learn. The standard environment would leave it up to users to select courses or lessons she likes to attend, but sometimes courses and lessons interconnection is too complex to understand or even hidden from the student. Therefore an artificial system based on clearly defined rules will be a perfect mentor existing inside the educational system. If courses are divided into isolated lessons, then it will be possible to skip certain, that the student doesn't need, or replace them with others bridging the gap in certain areas of student knowledge. This will ensure that all prerequisites for a successful understanding of next lessons are provided to the student.

Moreover, if the previous case will do some minor changes in a course, then more sophisticated engine could combine different lessons building a course on demand from available parts. Such artificial intelligence could easily produce non envisioned courses basing on personal characteristic either without authors supervision or minimizing it sufficiently. Obviously this is possibly only if the number of available lessons form a big enough database and those are well isolated.

The production of the delivered context basing on previous knowledge is not the only way to process a context available for the system. Another important property the system could monitor and measure is how well the student acquires the knowledge. For example the course could be a combination of explanation and practical tasks. The number of practical tasks could vary depending on how well previous were made in order not to slow down quick brains providing enough tasks to slow catchers.

The next important parameters the system could react on is the available time slots. The minimum requirement would be to preserve the position of a lecture where the student had to stop last time as the delivered content is now unique and so could be hard to reproduce exactly. Ideally the delivered portions should be dependent on availability of a person. For example she could define in advance time size of available time slots (for example 15 minutes each) and the system will deliver lessons of that size dividing those on topics and subtopics making the acquired knowledge tractable on the lowest grain (up to a lesson page).

The progress of learning could also be measured during the education and delivered in different level of explanation provided. The explanation could be short to keep the speed of moving over topics or could be very details if the person have no ideas at all about the explained concepts (or previous experience with those), so will require slowing down in this particular concept or a slow delivery of the knowledge in total.

The next important stage of the education process that can be a subject on context driven personalization is the examination. Sometimes results of the education process are a valuable achievement exposed to the business community. Those can be used to sort individuals basing on their skills and therefore the system should prevent the misuse of the education process by persons attempting to get higher marks than they actually are worth of. Therefore the study and examination process should be closely watched in order to identify the person taking the exam. There are hard and soft security system based on biometrical patterns that can be recorded communicated to a person. The basic idea here is to learn computer system to identify humans as they do in their everyday life distinguishing one person from another. Hard methods, like face recognition, require certain hardware to be used and therefore are not possible to be applied in many cases. Fortunately there are soft methods as well. One of the simplest "soft" biometrical identification methods is keystroke patterns, where the system monitors the keystroke typing dynamic, used symbols and words. It is well-known that writing/typing dynamics are assumed to be unique to a large degree among different people [6]. The same is known for the selection of words individuals use or keys on the keyboard (consider "delete" or "backslash" keys, which are nearly identical in functionality). Moreover the keystroke dynamic is not one factor to be measured but is a set of characteristics [7, 8, 9]:

1. Duration of the keystroke, hold time;
2. Latency between consecutive keystrokes;
3. Overall typing speed;
4. Habits by using additional keys on the keyboard like typing numbers via number pad part;
5. Frequency errors and backspace vs. delete key using;
   and some others

The pattern captured during tests and learning process can be reused during the examination phase and identify the student doing the exam. If there is any mismatching then the system should either have some kind person-to-person interaction possibility making possible to uniquely identify knowledge of the person or should have a concept of reliability of exam marks, which is delivered together with those, so interested parties could reevaluate those if needed.

Recent researches have reported that up to 65% of people over the middle age have certain disabilities and therefore could like to modify the user interface via which the content is delivered. This is another very important personal characteristic, which the system should react on. Notice that in many cases this is ignored in systems directed to students as quite a minor portion of young people suffer from any kind of disabilities. In the state wide system such ignorance is not possible any longer. The two elementary possibilities are required as a minimum. A possibility to increase the font size of the delivered content is one of those. Notice that although many browsers nowadays offer such functionality, not all web pages react on this correctly. The second one is possibility to deliver voice explanations. It will decrease the overall pressure on eyes for the education process participants and allow consuming the content by persons with disabilities. Moreover it could detach the education process from the computer as there are a lot of devices able to reproduce voice recording but not able to show web pages due the restricted screen size (or no screen at all) – consider as an example cell phones and mp3-players.

The age or disabilities is not the only reasons for slowing down the educational process. The overall environment students are located in could also promote or suppress the ability to understand information including the type of device the content is delivered too. Obviously even the sophisticated PDA device need restructuring delivered information as it doesn't have the same screen as a usual computer. Besides writing a text using the same keyboard at home or in the train is sufficiently different. Therefore the monitored keystroke patter used so far and property of the remote device could and should be used to adjust materials correspondingly.

B. *Protecting the content*

Nowadays, when knowledge becomes more and more important once again, many teachers are concerned about protect of their intellectual product (the course) from unauthorized copying. This becomes extremely important in the virtual environment, which is easy to access and even easier to copy and reproduce materials.

There are several ways to avoid this issue. First of all a copyright can be defined, which means a legal protection. Unfortunately it is not always good enough in virtual environments since the course can be used in other countries producing certain barriers for courts and laws. Secondly the author rights can be acquired by the state and deliberately unmonitored. The primary goal is to educate citizens and if it is achieved, then reuse of the course by anybody else can be ignored, as long as citizens do understand what courses they could take free of charge via the state system. The third and the last elementary protection mechanism is the language barrier, i.e. the national language used to build up the course prevents the same course to be transferred to countries with other languages solving a problem with countries where the law cannot be enforced to restore the course ownership.

The more sophisticated protection mechanism can be build basing on the acquired context during the process of education. The first problem that occurs quite often is to a) detect the node the course leaked out and b) prove who the real author of the course is. A so called "watermarking" could greatly help us resolving those two issues. It is possible to make the delivered context dynamic even if quite static files are delivered to the student by embedding into those invisible digital marks stating the owner of the content and the student it is delivered to. The process is basically similar to adding watermarks into banknotes except the fact that in the digital case we would like to keep them as hidden as possible in order to prevent removing of those. Notice that those marks are added right prior delivering the content to the student inside the virtual environment web server.

Besides, the earlier described personalization of the delivered context means that all lessons are divided into very small chunks which are delivered on demand basing on concrete person characteristics. In those small parts certain lessons can be omitted or replaced with others from completely other topics. All this could make the process of restoring the original content and reusing it to be very complex. It will require a lot of time and a certain level of knowledge making it even more complex that creating a course from scratch. Summarizing, the personalization divides the content into blocks making he course hard to acquire as a whole.

V. CONCLUSION

The virtual educational system described in the paper is an important component ensuring the social and financial stability of any state. The work-force productivity is one the most important factors defining the success of the state industry and innovation. The level of unemployment and professionalism (and consequently of salaries) defines the health of society including such as the level of crime or development of all state institutions. Therefore it is important to fund and developed such systems centrally by the state. The broad goal of the system makes it hard to build as it should serve educational needs of completely different people. Therefore the personalization of the delivered services becomes extremely important developing such systems. Fortunately the state either contains enough information about their citizens that can be consumed or the state system is trusted enough to enable context-driven collection of such data. Later the context can be converted into the concrete delivery of the education content to the concrete person during the concrete time-interval making the lesson unique and completely suitable to the available moment of life-time of the student. All this is designed to increase the knowledge acquisition level from the online system and close the gap between online education and teacher-student offline collaboration, which is usually much more intensive and so productive.

REFERENCES

[1] D. Mowery, R. Nelson, B. Sampat, and A. Ziedonis, "Ivory Tower and Industrial Innovation," *Stanford Business Books*, California: Stanford, 2004.

[2] K. Fischer, “The university as economic savior: as higher education replaces industry in some cities, colleges face both support and unrealistic expectations,” *Chronicle of Higher Education*, 2007.

[3] J. Youtie, and P. Shapira, “Building an innovation hub: A case study of the transformation of university roles in regional technological and economic development,” *Research policy*, vol. 37 (8), pp. 1188-1204, 2008.

[4] D. S. Siegel, D. A. Waldman, L. E. Atwater, and A. N. Link, “Toward a model of the effective transfer of scientific knowledge from academicians to practitioners: qualitative evidence from the commercialization of university technologies,” *J. of Engineering and Technology Management*, vol. 21(1-2), pp. 115-142, 2004.

[5] B. Readings, *Education / Teaching*, Harvard University Press, 1997.

[6] E. González-Agulla, E. Argones-Rúa, C. García-Mateo, and Ó W. M. Flórez, “Development and Implementation of a Biometric Verification System for E-learning Platforms,” *EDUTECH, Computer-Aided Design Meets Computer-Aided Learning, IFIP 18th World Computer Congress*, 2004, pp. 155-164.

[7] O. Guven, S. Akyokus, M. Uysal, and A. Guven, “Enhanced password authentication through keystroke typing characteristics,” *Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*, 2007, pp. 317-322.

[8] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14 (1), 2004, pp 4-19.

[9] A. K. Jain, L. Hong, and S. Pankanti, “Biometric Identification,” *Communications of ACM*, vol. 43 (2), 2000, pp. 91-98.

[10] D. Kumlander, “Soft Biometrical Students Identification Method for e-Learning,” *in Advances in Computer and Information Sciences and Engineering*, Springer, Netherlands, 2008, p. 114 – 118.

[11] P. Chang, A. Gul, "Towards Context-Aware Web Applications," 7th IFIP International Conference on Distributed Applications and Interoperable Systems (DAIS), Springer, LNCS 4531, 2007, pp. 239–252.

[12] W.Y. Lum, F.C.M. Lau, “A context-aware decision engine for content adaptation,” IEEE Pervasive Computing, vol. 1, no.3, 2002, pp. 41–49.

[13] J. Pascoe, “Adding Generic Contextual Capabilities to Wearable Computers,” Proc. 2nd Int’l Symp. Wearable Computers, IEEE CS Press, Los Alamitos, Calif., 1998, pp. 92–99.

Exploring Teachers' Beliefs about Digital Citizenship and Responsibility

Harrison Hao Yang
State University of New York at Oswego
Oswego, NY 13126, USA
harrison.yang@oswego.edu

Pinde Chen
South China Normal University
Guangzhou, 510631, China
pinde@scnu.edu.cn

***Abstract* – Promoting and modeling digital citizenship and responsibility can be a challenge to today's teachers. This paper presents a review of related studies connected with the digital citizenship movement and teachers' beliefs gleaned from the literature. It then reports a descriptive study which explores teachers' beliefs about and opinions on digital citizenship and responsibility.**

I. INTRODUCTION

Today's world has moved from the industrial age into the information age. The rapid technological change and proliferation of information resources are lineaments of our contemporary society. Multimedia on the Internet, telecommunications, wireless applications, electronic portable devices, social network software, Web 2.0, etc., are all radically redefining the way people obtain information and the way to teach and learn [1]. Findings of a 2007 national-wide survey study *Teens and Social Media* conducted by the Pew Internet and American Life Project show that some 93% of teens use the internet, and more of them than ever are treating it as a venue for social interaction – a place where they can share creations, tell stories, and interact with others. In addition, content creation by teenagers continues to grow, with 64% of online teenagers ages 12 to 17 engaging in at least one type of content creation, up from 57% of online teens in 2004 [2].

While the emerging technologies provide more opportunities and possibilities to enhance existing teaching and learning activities, there is considerable evidence that technology misuse and abuse is widespread and can be found inside and outside the school today: Web sites to intimidate or threaten students, downloading music illegally from the Internet, plagiarizing information using the Internet, using cellular phones during class time, and playing games on laptops or handhelds during class, etc [3] [4]. As a result, digital citizenship has been received the increased attention of many educators and has been considered as one of top priorities in numerous school districts. "Everyone – administrators, board members, teachers, parents and students – need to be involved in the dialogue about the appropriate use of technology" [3]. This paper reviews related studies connected with the digital citizenship movement and teachers' beliefs gleaned from the literature. It then provides a study which explores teachers' beliefs about digital citizenship and responsibility.

II. LITERATURE REVIEW

A. Digital Citizenship

The movement to address and characterize digital citizenship started in the UK, where educators have been working toward establishing protocols for good digital citizenship since the mid-1990s [5]. According to Digizen.org, owned and operated by London-based nonprofit Childnet International,

> Digital citizenship isn't just about recognizing and dealing with online hazards. It's about building safe spaces and communities, understanding how to manage personal information, and about being internet savvy - using your online presence to grow and shape your world in a safe, creative way, and inspiring others to do the same [6].

In the US, Ribble, Bailey, and Ross defined digital citizenship as "the norms of behavior with regard to technology use" [4]. They identified nine elements of digital citizenship: digital access, digital commerce, digital communication, digital literacy, digital etiquette, digital law, digital rights and responsibilities, digital health and wellness, and digital security. Exploring these elements has led an array of studies focusing on the implementation of good digital citizenship in schools, such as assessments, plans, strategies, activities, etc. [3] [4] [5] [7] [8] [9] [10] [11] [12]. As Ribble, Bailey, and Ross indicated "digital citizenship speaks to several levels of responsibility for technology. Some issues may be of more concern to technology leaders while others may be of more concern to teachers" [4]. The recent National Educational Technology Standards for Teachers 2008 (NETS·T 2008) prepared by the International Society for Technology in Education (ISTE) has attested the digital citizenship movement and specialized its Standard IV for teachers to "Promote and Model Digital Citizenship and Responsibility" [13]. Standard IV requests teachers "understand local and global societal issues and responsibilities in an evolving digital culture and

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_9, © Springer Science+Business Media B.V. 2010

exhibit legal and ethical behavior in their professional practices." In order to guide teachers meeting this standard, NETS·T 2008 Standard IV has four indicators covering different aspects of digital citizenship and responsibility:

- Indicator A: advocate, model, and teach safe, legal, and ethical use of digital information and technology, including respect for copyright, intellectual property, and the appropriate documentation of sources.
- Indicator B: address the diverse needs of all learners by using learner-centered strategies providing equitable access to appropriate digital tools and resources.
- Indicator C: promote and model digital etiquette and responsible social interactions related to the use of technology and information.
- Indicator D: develop and model cultural understanding and global awareness by engaging with colleagues and students of other cultures using digital-age communication and collaboration tools [13].

B. Teachers' Beliefs

In describing a working conception of beliefs, Bryan summarized "beliefs are psychological constructions that (a) include understandings, assumptions, images, or propositions that are felt to be true; (b) drive a person's actions and support decisions and judgments; …" [14]. It is important to realize that pre- and in-service teachers' opinions and briefs shape their actions. As Clark and Peterson pointed out:

> …the research shows that thinking plays an important part in teaching, and that the image of a teacher as a reflective professional…is not far-fetched. Teachers do plan in a rich variety of ways, and these plans have real consequences in the classroom. Teachers make decisions frequently (on every two minutes) during interactive teaching. Teachers do have theories and belief systems that influence their perceptions, plan, and actions [15].

Similarly, Cuban stated, "The knowledge, beliefs, and attitudes that teachers have shape what they choose to do in their classrooms and explain the core of instructional practices that have endured over time" [16]. Pajares found there was a "strong relationship between teachers' educational beliefs and their planning, instructional decisions, and classroom practices" and that "research on the entering beliefs of pre-service teachers would provide teacher educators with important information to help determine curricula and program direction and interpretation of knowledge and subsequent teaching behavior" [17].

In sum, the literature suggests that it is critical to investigate the pre- and in-service teachers' beliefs about technology innovation and integration. Specifically, the pre- and in-service teachers' beliefs and opinions indicate their willingness to adopt a new teaching behavior such as promoting and modeling digital citizenship and responsibility in their professional practices. Although previous studies have provided a variety of practical ways on the implementation of good digital citizenship in schools, such as assessments, plans, strategies, activities, etc. in general, few studies have attempted to investigate teachers' beliefs about digital citizenship and responsibility in particular.

III. The Study

The purpose of this study was to examine teachers' beliefs about and opinions on NETS·T 2008 Standard IV - *Promote and Model Digital Citizenship and Responsibility*. The following questions guided this study:

1. What were teachers' beliefs about and opinions on each indicator of NETS·T 2008 Standard IV?
2. How could teachers adopt and implement the four indicators of NETS·T 2008 Standard IV into their teaching practices?

A. Participants and the Course

The participants of this study came from four sections of students (n = 87) who were enrolled in the asynchronous online course entitled *Multimedia and Internet for Educators*, offered at a university in the northeastern region of the United States during the winter, spring, summer, and fall terms in 2009. Most of the participants were part-time in-service teachers located in the same geographical area, who were pursuing graduate level education programs in content areas of literacy, biology, chemistry, earth science, mathematics, social studies, technology, etc.

The course *Multimedia and Internet for Educators* provides an introduction and guide to pre- and in-service teachers to current and emerging technology. It intends to help PK-12 educators not only to use multimedia and Internet resources in their own educational endeavors, but also integrate technology into their work environments as teachers. Since spring 2001, this course has been joined in the State University of New York Learning Network (SLN) as one of its asynchronous learning network courses offered to both on and off-campus students and is an elective graduate course. In each module of this course, students were required not only to complete and submit the assignment, but also to collaborate on specific learning activities required during a certain time frame: (a) reading each "Mini-Lecture" and supplemental materials, (b) responding to the questions in each Mini-Lecture at the "Online Discussion" section, (c) responding publicly to some or all of the questions submitted by classmates, and (d) replying to classmates who responded to the question. Asynchronous messages posted by the participants were threaded and could be sorted by author or by date for the participants' viewing and reviewing preference.

B. Data Collection

To investigate teachers' beliefs about and then to promote their understanding of digital citizenship and responsibility, in the discussion area of *Module 1 - Introduction of Multimedia and Internet* the participants were asked to provide the written responses to the following statement:

> *People are excited by the power and popularity of the Internet, in your opinion, what are the ethical/social/human issues we should consider when we implement digital technologies, especially the Internet-based technologies into today's education?*

As indicated in Table 1, a total of 299 responses from participants were collected and analyzed by the instructor of the course, who is the first author of this study as well.

TABLE 1
NUMBERS OF PARTICIPANTS AND RESPONSES

	Participants	Responses
Winter 09	26	92
Spring 09	20	69
Summer 09	21	91
Fall 09	20	47
Total	87	299

C. Findings

Among 299 responses, while most of them addressed issues ranged over two or more indicators of NETS·T 2008 Standard IV, some of them focused on one indicator of NETS·T 2008 Standard IV. Table 2 listed numbers and percentages of responses that pertained to each indicator.

TABLE 2
DISTRIBUTION OF PARTICIPANTS' RESPONSES

	Frequency	Percent
Indicator A	210	70.2
Indicator B	44	14.7
Indicator C	125	41.8
Indicator D	107	35.8

First of all, the vast majority of responses (70.2%) were falling into the Indicator A – "*advocate, model, and teach safe, legal, and ethical use of digital information and technology, including respect for copyright, intellectual property, and the appropriate documentation of sources*." The following online discussions in fall 2009, which involved three participants exchanging their ideas on the plagiarism, were typical examples:

- [The initial discussion on the plagiarism posted by Student A-Fall]
  Researching for papers the internet makes the availability to plagiarism a lot easier for students. Students even have the option to purchase papers of the internet. The authenticity of individual work may be more difficult to determine with the access of the internet.
- [Student B-Fall responded to Student A-Fall's initial post]
  I agree that it is much more convenient and easy for students to plagiarize while using the World Wide Web (sometimes without even knowing it!). Due to their early start to using the internet and their lack of knowledge about plagiarism this could definitely be a problem that might occur when using the internet in the classroom. I think as educators we need to address this problem, educate our students about plagiarism and teach them how to use the information correctly. If a teacher chooses to use the internet for a research project I think it is essential to address this issue in the very beginning and spend time discussing how to complete a works cited to use the information legally.
- [Student A-Fall responded to Student B-Fall's response]
  I think that if a teacher chooses to use the internet during any type of project it is essential that the students learn it is not appropriate to copy someone else's work. On top of that I think it is the teacher's job to double check each student's work and make sure they haven't done so. I remember from student teaching placement when I gave an essay assignment and had a student plagiarize I took them aside and asked why they did so. They told me they had never gotten caught before so they thought they could get away with it. They understood they did something wrong but had never gotten caught before. The reinforcement of being caught would help prevent it in the future.
- [Student C-Fall responded to Student B-Fall's response]
  Whenever students are assigned to write an essay in which they need to do research I feel the teacher should always give a mini lecture on the importance of citing work to avoid plagiarism as well as making students aware of the consequences they will face if they are caught in the act of plagiarism. Students should also be reminded of how to determine if a source is reliable and creditable as well as review on how to site works (or a handout) if necessary. My content area is business but I always discussed citations, references, and plagiarism with my students as it is important in every subject.

Secondly, although there were fewer responses (14.7%) on the Indicator B - "*address the diverse needs of all learners by using learner-centered strategies providing equitable access to appropriate digital tools and resources*," participants thoroughly discussed this important area regarding the implementation for their teaching practices. As were evident in the following participants' discussion in spring 2009:

- [The initial discussion on the digital access posted by Student A-Spring]
  The major social issue that I have been faced with is the fact that many of our students do not have access at home and it is important to be fair when giving assignments which require this. In fact, I have gone so far as to meet students at the local library to help them with a project that I assigned, so that they would have the time outside of class, but my support, since they may not be as familiar with the technology as their peers.
- [Student B-Spring responded to Student A-Spring's initial post]
  I applaud you for going to the local libraries to help your students. Why don't they use the computers in school? There are a number of students who don't have a computer at home do you have alternative version of the assignment for them?
  I do not think of lack of computers at home as a social issues. Good thinking. I hope this is the reason why

some schools are developing laptop programs for their students.

- [Student A-Spring responded to Student B-Spring' response]
  Usually it's just one or two students who don't have the access and it's for those kids who I will go above and beyond - although I could provide an alternate assessment, it is important to me that they have the authentic experience like their peers.
- [Student C-Spring responded to Student B-Spring's response]
  My school had the laptop program (it was a predictable disaster, and no longer is offered), and only the students who were leasing one of the laptops ($$) got to take them home. The students who could not afford to lease them only had access to them during the day. They had to return them before going home. So, the disadvantaged students who probably did not have a computer at home to begin with still couldn't have access at home because they couldn't afford to lease the laptops and take them home. From the financial standpoint, it makes sense. The district can't afford to have kids "lose" $800 laptops for which there will be no financial recovery.
- [Student D-Spring responded to Student A-Spring' initial post]
  As a teacher I am also facing with the same social issue in my own classroom. It is great that students can access their homework from the internet or interact with each other in a math game, but only for the students that have a computer at home. Many of my students do have a computer, but a few don't. The ones that do have a computer don't always have internet access at home. I choose not to really assign any homework that requires using the internet for this reason. We use it within our classroom.
- [Student E-Spring responded to Student D-Spring' response]
  I agree, and only require use of computers when I am supplying them in my classroom via laptop carts. It surprises me that so many people fail to consider students lacking access to a computer when creating assignments.

Thirdly, approximately 42% of the responses addressed various aspects on the Indicator C – "*promote and model digital etiquette and responsible social interactions related to the use of technology and information*." The following online discussions in winter 2009, which involved three participants exchanging their ideas on the cyber-bulling, were typical examples regarding one of new-fangled and concrete issues related to Indicator C:

- [The initial discussion on the cyber-bully posted by Student A-Winter]
  Bullying has always been an issue in schools among students. Now they are able to do it behind computer monitors and cell phones. Some of this is being done right from computers and cell phones at school. It's their new form of bathroom graffiti and harassment. Teachers can't witness this cyber-bullying and parents don't even know it's going on most of the time. I think this gives the bullies an even bigger sense of power. They are unseen, and sometimes even nameless, except for their screen name, which maybe is known or unknown. When students are given computer usage time, it needs to be carefully monitored by the network administrator and the supervising teacher. A students' history can be tracked, but it is difficult for a teacher to monitor 20 students' activity when all it takes is a click of the mouse to close a window you're not supposed to be in. What are some effective ways to monitor computer usage by a classroom of students? How can educators try to eliminate cyber-bulling in their schools?
- [Student B-Winter responded to Student A-Winter's initial post]
  As I think about your question about how educators can try to eliminate cyber-bulling in our schools it just brings up such a difficult task. It makes me think of just regular bullying and how much of it occurs outside of school. How do we combat that? I've talked with parents before whose children are on the social networks and they said from their standpoint, it's just lots of monitoring. If anything, it seems like cyber-bullying is easier to monitor because they leave a trail of everything they are doing.
- [Student C-Winter responded to Student A-Winter' initial post]
  Great post and great questions! I found this article *Tools Help Families Combat Cyber Bullying with Alerts, Tips* about a tool called "CyberBully Alert" and thought the information was interesting to share - "One such tool, recently unveiled by Vanden, allows children to instantly notify selected adults when they are bullied or harassed online. CyberBully Alert (http://www.cyberbullyalert.com/) also documents the threatening message by saving a screen shot of the child's computer when the child triggers an alert. The Web-based solution is compatible with most computers and Web browsers. It requires a single download and registration. Parents then select which e-mail addresses and phone numbers will receive the alerts. Alerts are sent through e-mail and text message when a child clicks on the CyberBully Alert icon… The creators of CyberBully Alert also offer a School Incentive Program" (cited from Information Week, October 1, 2008).

Finally, there were approximately 36% of the responses falling into the Indicator D – "*develop and model cultural understanding and global awareness by engaging with colleagues and students of other cultures using digital-age communication and collaboration tools*." Participants reported their deep awareness of this indicator. As were evident in the following participants' discussion in summer 2009:

- [The initial discussion posted by Student A-Summer]

My concern is that when students can be a part of the "social clique" created with technology, how will that affect their ability to be social in the real world? When someone is either at the keyboard or with headsets on, will their real world collapse into email, YouTube and iPod? Will families still have the same connections, when each as an individual has the ability to create their Second Life avatar, design their own room and hang out with people who are only like them? Are we building communities for our children that are physical and in real time? I realize these are questions, but when our students are being prepared for a high technology workplace, are their social and moral obligations to each other as community citizens being addressed? I also want to note that as our youth are exposed the ethical and social differences of other countries, other religions and the range of human behavior throughout the world, will our culture be affected? That is where educators have the moral obligation of understanding how we are exposed to differences and how they are discussed can have impact our personal lives and future choices.

- [Student B-Summer responded to Student A-Summer's initial post]
  The issues you have raised should be a concern for everyone. More and more people are not communicating face to face. I think the implications that this can have on society can be detrimental to how we all communicate with each other. I sometimes find myself texting to a friend or family member when I could easily call them and verbally say what I needed to. Progressively, since the continuing line of new advancements in technology, the idea of community has declined in society. It's scary to think that more and more people have replaced the ideas of community with thoughts concerning only themselves. As educators I think it is important that we continually stress the importance of community and help students to understand the value of working together. By exposing students to the benefits of community it will help them understand that in order to thrive, community cooperation is essential. Also, I think that our culture will undoubtedly be affected. Cultural diffusion does not only take place through means on face to face interaction. With the internet students can view videos of virtually any society that they wish. Students can find information on remote villages in the Amazon rainforest, something few students of past generations could do. I do believe that this exposure to other cultures will be beneficial to students. Educators can utilize this resource to help teach students to understand other cultures and societies, working to break stereotypes and prejudices. I hope that affects on our culture will be beneficial; however educators need to understand their importance to this cultural diffusion.
- [Student C-Summer responded to Student A-Summer's response]
  This is an interesting thought because there are many communication resources that students can use. There is a fine line between replacing technology-based communication with face to face communication. Personally, I believe the technology-based communication that we have today is helping our society stay in touch. Although it is not face to face communication, it does not mean that people are replacing technology with going out and meeting people. This technological communication surge is simply helping stay in touch with people on a 24-7 basis.

IV. CONCLUSION

The findings of this study may produce useful information that will enable administrators, program planners, and teachers to better understand teachers' beliefs about digital citizenship and related responsibilities. As Ribble and Bailey pointed, "The endless list of misuse and abuse includes hacking into school servers, using e-mail to intimidate or threaten students, illegally downloading music, plagiarizing information from the Internet, using cellular phones during class time, accessing pornographic Web sites, and playing video games during class. Therefore, if you are using technology in your district, you must begin to deal with digital citizenship in a significant way" [10]. To capture the improvements in digital citizenship and responsibility, teachers must be prepared – through sharing and expressing their beliefs and opinions – to learn what digital citizenship is and what it can do. Indeed, there have been some of valuable resources and activities related to digital citizenship existing for teachers and educators working on curriculum. For example, for teachers to teach Internet safety, some of lessons and activities including electronic simulations are available online [18]:

- Cybersmart (http://www.cybersmart.gov.au/) is developed by the Australian Communications and Media Authority, and "provides activities, resources and practical advice to help young kids, kids, teens and parents safely enjoy the online world" [19].
- i-SAFE (http://www.isafe.org/) is endorsed by the U.S. Congress, and "incorporates classroom curriculum with dynamic community outreach to empower students, teachers, parents, law enforcement, and concerned adults to make the Internet a safer place" [20].
- NetSmartz Workshop (http://www.netsmartz.org) "is an interactive, educational safety resource from the National Center for Missing & Exploited Children® (NCMEC) and Boys & Girls Clubs of America (BGCA) for children aged 5 to 17, parents, guardians, educators, and law enforcement that uses age-appropriate, 3-D activities to teach children how to stay safer on the Internet" [21].
- ID the Creep (http://www.idthecreep.com) is an online game sponsored by the National Center for

Missing & Exploited Children and The Ad Council for young students to practice identifying pedophiles online.

If the strong beliefs can be built, then better teaching behavior and learning activities can be developed and implemented. For this reason, beliefs about digital citizenship and responsibility among teachers may be an important consideration in the process of incorporating NETS·T into real-world teaching, with options for more concrete actions for teachers.

REFERENCES

[1] S. Wang and H. Yang, "Editorial," *International Journal of Technology in Teaching and Learning*, vol. 3 issue 2, 2007.

[2] A. Lenhart, M. Madden, A. R. Macgill, and A. Smith, "Teens and social media: The use of social media gains a greater foothold in teen life as they embrace the conversational nature of interactive online media," *Pew Internet & American Life Project*, 2007. <http://www.pewinternet.org/~/media//Files/Reports/2007/PIP_Teens_Social_Media_Final.pdf.pdf >

[3] M. S. Ribble, "Point of view on technology drivers' licenses," *District Administrator Magazine*, 2004. < http://www.digitalcitizenship.net/uploads/TechDL.pdf>

[4] M. S. Ribble, G. D. Bailey, and T. W. Ross, "Digital citizenship: Addressing appropriate technology behavior," *Learning & Leading with Technology*, vol. 32 issue 1, pp. 7-11, 2004.

[5] M. Villano, "Text unto others…as you would have them text unto you," *T. H. E. Journal*, September 2008. <http://thejournal.com/articles/2008/09/01/text-unto-others-as-you-would-have-them-text-unto-you.aspx>

[6] *What makes you such a good digizen?* < http://www.digizen.org/>

[7] J. Sturgeon, "Raising good digital citizens: Teachers are developing new ways for students to follow good internet and media use practices," *Cable in Classroom*, November 2008.

[8] M. S. Ribble and G. D. Bailey, "Digital citizenship: Focus questions for implementation," *Learning & Leading with Technology*, vol. 32 issue 2, pp. 13-15, 2004.

[9] M. S. Ribble and G. D. Bailey, "Developing ethical direction," *Learning & Leading with Technology*, vol. 32 issue 7, pp. 36-37, 2005.

[10] M. S. Ribble and G. D. Bailey, "Monitoring technology misuse & abuse: A five-step plan for creating a digital citizenship program in your school," *T. H. E. Journal*, August 2004. < http://www.digitalcitizenship.net/uploads/T.H.E.JournalArticle.pdf>

[11] M. S. Ribble and G. D. Bailey, "Is digital citizenship a problem in your school?" *Technology Turning Points*, December 2005. <http://www.digitalcitizenship.net/uploads/TechnologyTurningPointsTeacherTalk.pdf>

[12] M. S. Ribble and G. D. Bailey, *Digital citizenship in schools*, ISTE, 2007.

[13] International Society for Technology in Education, *National education technology standards for teachers*, 2008. <http://www.iste.org/Content/NavigationMenu/NETS/ForTeachers/2008Standards/NETS_for_Teachers_2008.htm>

[14] L. A. Bryan, "Nestedness of beliefs: Examining a prospective elementary teacher's belief system about science teaching and learning," *Journal of Research in Science Teaching*, vol. 40 issue 9, pp. 835-868, 2003.

[15] C. M. Clark and P. L. Peterson, "Teacher's thought process," in M. C. Wittrock (ed.). *Handbook of research on teaching*, New York: McMillan Publication, 1986.

[16] L. Cuban, "Computers meet classroom: Classroom wins," *Teachers College Record*, vol. 95 issue 2, pp. 185–210, 1993.

[17] F. M. Pajares, F. M., "Teachers' beliefs and educational research: Cleaning up a messy construct," *Review of Educational Research*, vol. 62 issue 3, pp. 307-332, 1992.

[18] A. McAnear, "Digital citizenship," *Learning & Leading with Technology*, 2007. <http://www.iste.org/Content/NavigationMenu/Publications/LL/LLIssues/Volume_34_2006_2007_/March_6_/34604m.pdf>

[19] < http://www.cybersmart.gov.au/>

[20] < http://www.isafe.org/>

[21] < http://www.netsmartz.org/overview/aboutus.htm>

Faculty Driven Assessment of Critical Thinking: National Dissemination of the CAT Instrument

Barry Stein, Ada Haynes, Michael Redding, Kevin Harris, Mark Tylka, and Elizabeth Lisic
Tennessee Technological University

***Abstract*- This paper reports the latest findings of a project to nationally disseminate the CAT© instrument, a unique interdisciplinary assessment tool for evaluating students' critical thinking skills. Tennessee Technological University originally partnered with six other institutions across the U.S. to evaluate and refine the CAT instrument beginning in 2004. Through these efforts a test with high face validity, high construct validity, high reliability, and that is culturally fair was developed In the current national dissemination project, we are collaborating with over 40 institutions across the country to disseminate this unique instrument and support the creation of assessment based efforts to improve student learning. These dissemination efforts involved training representatives from other institutions to lead scoring workshops on their own campuses as part of a broader effort to improve student learning. A variety of findings indicate these dissemination efforts are successful. National user norms that are being collected from this project indicate that largest gains in undergraduate critical thinking appear to occur in the junior and senior years. [1]**

INTRODUCTION

In our increasingly technological and information driven society, the ability to think critically is a cornerstone to both workplace development and effective educational programs. The importance of critical thinking skills in higher education has become increasingly apparent. Accrediting agencies such as Accreditation Board of Engineering and Technology (ABET) recognize the need for higher order thinking skills and real world problem solving in their accreditation standards [1]. Faculty also recognize the importance of critical thinking skills. According to Derek Bok [2], over ninety percent of the faculty in the U.S. feel that critical thinking is the most important goal of an undergraduate education. Despite the recognized importance of critical thinking skills, it is difficult to find effective tools for assessing these skills. Many of the assessment tools are plagued by problems related to validity, reliability, and cultural fairness [3]. According to Bransford, Brown, and Cocking [4] "a challenge for the learning sciences is to provide a theoretical framework that links assessment practices to learning theory." Pellegrino, Chudowsky and Glaser also note that assessment instruments need to be developed based upon principles of learning and cognition [5].

[1] Partial support for this work was provided by the National Science Foundation's CCLI Program under grants 0404911 and 0717654. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Tennessee Technological University (TTU) has been engaged in an extended effort during the last 9 years to develop and refine an instrument to assess critical thinking that overcomes many of the weaknesses of other existing tools. Preeminent theoreticians and educators in the area of learning sciences and assessment participated in the project. Unlike many other available assessment tools, the Critical thinking Assessment Test (CAT) instrument uses short answer essay responses to assess critical thinking. Although essay questions can be harder to score than multiple choice questions, they provide a better understanding of students' thought processes and ability to think critically and creatively when confronted with real world problems [6]. In addition, the CAT instrument is unique in that it utilizes a campus's own faculty to evaluate student responses. Ewell [7] has noted that one problem with many existing assessments is that "we let assessment get excessively distanced from the day-to-day business of teaching and learning." The CAT instrument allows faculty to directly observe students' critical thinking and understand their students' weaknesses. This activity helps motivate faculty to consider changes in pedagogy that might improve students' critical thinking skills [8]. This becomes increasingly important as accrediting agencies such as ABET increase their focus on efforts to improve students' critical thinking [1].

TTU's approach to developing the CAT instrument has been to involve faculty in the identification of a core set of skills they believe to be an important part of critical thinking across many disciplines (see Table 1). TTU worked with a diverse group of institutions nationwide (University of Texas, University of Colorado, University of Washington, University of

TABLE 1
SKILL AREAS ASSESSED BY THE CAT INSTRUMENT

Evaluating Information
Separate factual information from inferences.
Interpret numerical relationships in graphs.
Understand the limitations of correlational data.
Evaluate evidence and identify inappropriate conclusions.
Creative Thinking
Identify alternative interpretations for data or observations.
Identify new information that might support or contradict a hypothsis.
Explain how new information can change a problem.
Learning and Problem Solving
Separate relevant from irrelevant information.
Integrate information to solve problems.
Learn and apply new information.
Use mathematical skills to solve real-world problems.
Communication
Communicate ideas effectively.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_10, © Springer Science+Business Media B.V. 2010

Hawaii, University of Southern Maine, and Howard University) to further develop, test, and refine this instrument. The results of that project revealed the CAT instrument had high face validity when evaluated by a broad spectrum of faculty across the U.S. in STEM and non-STEM disciplines, had good criterion validity when compared to other instruments that measure critical thinking and intellectual performance, had good scoring reliability, good test/re-test reliability, and good construct validity using expert evaluation in the area of learning sciences [8].

Student response to the CAT instrument has also been very positive. Many students indicate appreciation for the real world problem solving tasks that make up the test [9]. Using questions that are interesting and engaging helps motivate students to perform their best.

Performance on the CAT instrument is significantly correlated with other measures of student performance including cumulative college grade-point average (r = 0.295), entering SAT or ACT scores (r = 0.528, and r = 0.560, respectively), and performance on several other tests that are thought to measure certain skills associated with critical thinking such as the California Critical Thinking Skills Test (r = 0.645) and the CAAP Critical Thinking Module (r = 0.691). These correlations provide support for the criterion validity of the CAT instrument. However, the magnitude of the correlations also indicate that the CAT instrument measures something different than the other assessment tools [9]. Performance on the CAT instrument has also been found to significantly correlate with several key questions on the National Survey of Student Engagement (NSSE). One NSSE question that is significantly negatively correlated with students' performance on the CAT instrument is the extent to which students think their courses emphasized rote retention [10]. In contrast, CAT performance is significantly positively correlated with NSSE questions that reflect the extent to which students' think their college experience emphasized critical thinking and real-world problem solving.

Extensive work has been done to refine the scoring process for the CAT instrument to improve scoring accuracy and reliability. For example, the scoring process uses multiple scorers for each question and uses a detailed scoring guide that has been refined with extensive faculty input to ensure scoring consistency. Although scoring accuracy is typically a problem for essay type tests, the scoring process for the CAT instrument overcomes many of the problems typically associated with scoring essay tests.

The major focus of the new work reported in this paper involves efforts to disseminate the instrument to a broad range of institutions across the U.S. These efforts involved training representatives from other institutions to lead scoring workshops on their own campuses as part of a broader effort to improve student learning.

METHOD

Funding from the National Science Foundation supported collaboration with 20 institutions across the country to administer the CAT instrument to their students and to conduct scoring workshops with their own faculty to score their students' responses on the CAT instrument. Since institutional interest in the CAT instrument greatly exceeded that level of participation supported by NSF, we created an alternative method of institutional participation that was self-supporting. We are currently collaborating with over 40 institutions across the country. Grant funded participants were selected to insure diversity in mission, institutional characteristics, ethnicity/racial demographics, and geographic location. The collaborating institutions range from community colleges to large research institutions and include private, as well as, publicly funded institutions.

Materials

A variety of training materials were developed to support the training of new scoring workshop leaders and to help those leaders train their own faculty to effectively score the CAT instrument on their respective campuses. These materials included a narrated overview of the development process for the CAT instrument that streams from our website (www.CriticalThinkingTest.org), a detailed training manual that covers issues from sampling and obtaining institutional review board approval to administering the test, recruiting faculty scorers, conducting the scoring workshop, and interpreting institutional results. A detailed scoring guide and multi-media computer based training modules were also developed to explain question scoring.

Train-the-Trainer Workshops

Participating institutions sent two to three representatives to a two-day training workshop where they received extensive training in scoring the CAT instrument and preparation to lead a scoring workshop on their own campus. The workshop also covered other topics related to understanding and using different assessment designs, sampling, and using the scoring session as an opportunity for faculty development. One of the advantages of the CAT instrument is that it gives faculty a good understanding of student weaknesses. Coupling this experience with a discussion of ways to improve student learning creates an effective faculty development experience that can lead to real improvements in instruction and student learning.

RESULTS

The effectiveness of the train-the-trainer workshops were evaluated in a number of ways including; participant surveys, direct observations by our external evaluator, follow-up onsite observations of campus scoring sessions, and scoring accuracy checks of tests scored by each institution.

Training Workshop Surveys

At the conclusion of each of the training workshops participants were asked to complete an anonymous survey to evaluate various components of the workshop training. The survey used a 6 point Likert type scale: 6 – Strongly Agree, 5 –

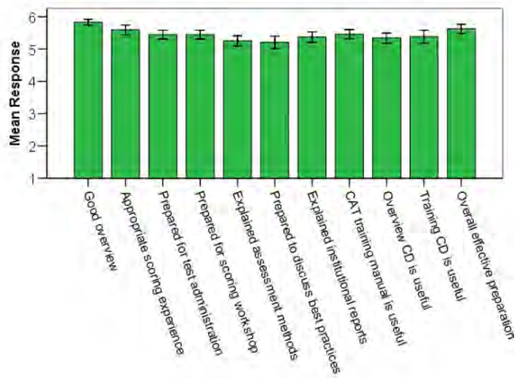


Fig. 1
Average Training Workshop Evaluation Responses (with 95% CI)

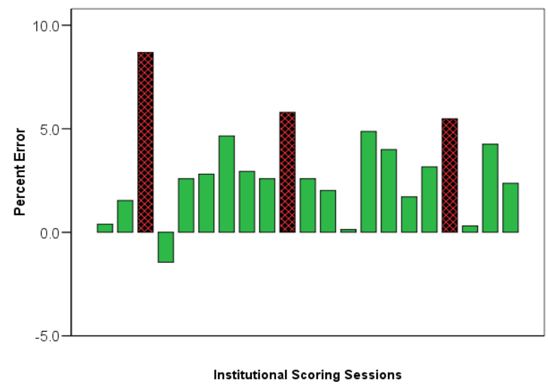


Fig. 2
Results of Scoring Accuracy Checks Across Institutions as Percent Error

Agree, 4 – Slightly Agree, 3 – Slightly Disagree, 2 – Disagree, 1 – Strongly Disagree. Eighty-five participants were included in this analysis.

The mean response on each survey item is shown in Figure 1 (error bars indicate 95% confidence interval). The results indicate that the workshop evaluations have been quite positive across the topics surveyed.

Qualitative Evaluation of Training Workshops

Observations by our external evaluators and project staff are used formatively to continually improve the training workshops. For example, we noted early in the project that participants in the training workshops may find it difficult to process information from too many different sources (the workshop leader, the multimedia training modules, and the scoring guide). This led to changes in how the multi-media training module was used in the workshop. It is now used during the second day as a resource for explaining scoring in situations where scorers may need further explanation. Many other changes were also made in response to qualitative evaluations (e.g., extending the discussion of sampling methods and exploring methods for obtaining representative samples at different institutions, and encouraging participants to develop discipline specific analog learning activities that correspond to questions on the CAT instrument).

Onsite Evaluation of Scoring Workshops

Observations by our external evaluators and project staff who have attended scoring workshops at other institutions have been particularly useful. These observations have confirmed that procedures outlined in our training manual must be carefully adhered to in order to efficiently and accurately score tests. Institutions that have deviated in important ways from those guidelines experienced numerous problems. These observations have led to increased emphasis and training support to explain various procedures associated with test scoring.

Scoring Accuracy Checks

One of the most important pieces of evaluative information about the success of this national dissemination model comes from checking the accuracy of scoring performed by each participating institution. Experienced scorers at our institution rescore randomly selected tests from each scoring session at other institutions. The sample includes about 15% to 20% of the tests scored by each institution. Error rates below 5% are considered acceptable. Thus far, overall accuracy has been very good. Figure 2 shows the overall test scoring error found across 21 institutional scoring sessions that have undergone accuracy checks. About 1 in 7 institutions has been found to deviate more than 5% from our experienced scorers. All institutions receive a scoring accuracy report together with a question by question analysis of accuracy, and suggestions for improving scoring accuracy in subsequent sessions. Recommendations are given to institutions that deviate more than 5% on how to appropriately adjust their scores for comparison to other institutions.

National Norms

Another goal of the current project is to begin assembling national norms for various populations of students including community college students and students at four-year institutions. Although many institutions will track their progress by

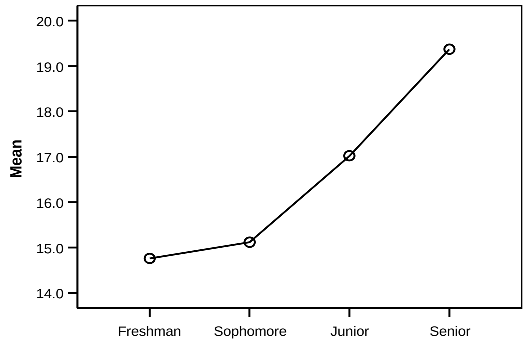


Fig. 3
Mean Scores on the CAT Instrument for Students at 4-year Institutions

comparison to their own test scores over a period of time, others have requested information that would allow comparison to larger populations. Figure 3 shows the mean score on the CAT instrument for approximately 3000 students of different class standings at four-year institutions across the country. The largest gains in critical thinking appear to occur in the junior and senior years of undergraduate education. The average senior score was about 51% of the maximum possible score on the CAT instrument (maximum score = 38 points). These scores indicate there is considerable room for improvement. The fact that no student or has obtained a perfect score or a score of zero at any institution suggests the absence of a ceiling or floor effect.

CONCLUSION

There have been several significant outcomes of the current national dissemination project. We have found that our dissemination model for training representatives at other institutions to use the CAT instrument at regional train-the-trainer workshops is successful. This finding is supported by information from participant surveys, onsite observations, and scoring accuracy checks. We do, however, continue to modify our training methods and supporting materials to improve effective dissemination.

We have found that many institutions are interested in finding ways to engage faculty in quality improvement efforts. Faculty involvement in the scoring of student essay exams greatly facilitates this process. We are also expanding our efforts to provide links to institutional and grant funded projects that have positively impacted student performance on the CAT instrument. This type of information is just beginning to emerge from a broad range of institutional initiatives. We are also collaborating with several other NSF funded projects that are using the CAT instrument to help identify potential student gains in critical thinking and real-world problem solving that may result from innovative educational pedagogies.

The CAT instrument is one of the few interdisciplinary assessment instruments available that also provides an opportunity for faculty development. By participating in the scoring process, faculty become aware of their students' weaknesses and can begin to explore modifications in teaching methods that might address these weaknesses. Many of the participants in our regional training workshops feel that the CAT instrument is particularly useful because it helps faculty understand how they can develop discipline specific assessments that encourage students to think critically instead of just encouraging rote retention. This becomes increasingly important as accrediting agencies such as the Accreditation Board of Engineering and Technology (ABET) increase their focus on efforts to improve student learning.

REFERENCES

[1] Accreditation Board of Engineering and Technology, *Criteria for Accrediting Engineering Programs: Effective for Evaluations During the 2006-2007 Accreditation Cycle.* Baltimore, MD: Accreditation Board of Engineering and Technology, 2005

[2] D. Bok, *Our Underachieving Colleges: A candid look at how much students learn and why they should be learning more.* Princeton: Princeton University Press, 2006.

[3] U.S. Department of Education, National Center for Education Statistics. *The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing,* NCES 2000—172, prepared by T. Dary Erwin for the Council of the National Postsecondary Education Cooperative Student Outcomes Pilot Working Group: Cognitive and Intellectual Development. Washington, DC: U.S. Government Printing Office, 2000.

[4] J. Bransford, A. Brown, and R. Cocking, (Eds.) *How People Learn: Brain, Mind, Experience, and School.* Washington, D.C.: National Academy Press. 2000.

[5] J. Pellegrino, N. Chudowski, and R. Glaser (Eds.) *Knowing What Students Know.* Washington, D.C., 2001.

[6] B. Stein, A. Haynes, and M. Redding. "Project CAT: Assessing Critical Thinking Skills," in *Proceedings of the National STEM Assessment Conference,* D. Deeds and B. Callen , Eds. NSF, 2007.

[7] P. Ewell. "The examined life: Assessment and the ends of general education," Association of American Colleges & Universities General Education and Assessment Conference, 2004.

[8] B. Stein, A. Haynes, M. Redding, T. Ennis, and M. Cecil. "Assessing Critical Thinking in STEM and Beyond," in *Innovations in E-learning, Instruction Technology, Assessment, and Engineering Education*, M. Iskander, Ed. Springer, 2007, pp.79-82.

[9] B. Stein, A. Haynes, and K. Harris. "Assessing and Improving Critical Thinking Using the CAT Instrument," *ABET Best Assessment Processes Symposium XI*. Indianapolis, IN: Accreditation Board of Engineering and Technology, 2009.

[10] M. Cecil, "The Relationship Between Student Responses on the National Survey of Student Engagement (NSSE) and Performance of the Critical-Thinking Assessment Test (CAT)," Dissertation, Tennessee Technological University, 2006.

Enhancing the Instruction of Introductory Electric Circuit Courses using MATLAB

Thomas Yang, *Member, IEEE,* Ilteris Demirkiran, *Member, IEEE,* Jianhua Liu, *Member, IEEE*, Albert Helfrick, *Senior Member, IEEE*, David Pedersen, and Christopher Grant

***Abstract*—This paper discusses the implementation and assessment of the integration of MATLAB usage into the instruction of an introductory circuit analysis course at Embry-Riddle Aeronautical University (ERAU). This teaching practice seeks to enhance student learning by delivering relevant course materials using MATLAB's graphical representation and problem solving capabilities. Also, students learn to plot voltage/ current waveforms and other important graphs via simple MATLAB programs. The effectiveness of the teaching practice is assessed through assignments and a survey. Assessment data indicates that integration of MATLAB into course instruction promotes learning and enhances students' computer skills. In addition, students highly appreciate the effectiveness of this teaching practice.**

***Index Terms*— MATLAB, circuit analysis, assessment, engineering education**

I. Introduction

Widely considered as the "language of technical computing" [1], MATLAB is used extensively in the scientific and engineering communities for research and development. With its universal approach to use the matrix as a fundamental data structure, its large library of pre-made functions, and readily available toolboxes, Matlab finds application in all fields of engineering [2]. Hence, the ability to develop solutions to technical problems using MATLAB is becoming increasingly important for graduates of engineering schools. MATLAB allows students to focus on course work and applications rather than on programming details. Writing MATLAB codes to perform numerical calculations only requires a fraction of the time it would take to write a program in a lower level language. MATLAB helps engineers better understand and apply concepts in applications [3]. Powerful graphical capabilities and simple grammar structures make MATLAB ideal for assisting classroom instruction. It is also capable of producing interactive simulation. This can enhance the quality of presentation with graphical simulations [4].

Although there have been several studies on using MATLAB in electrical engineering curriculum [5-10], we believe there has not been enough emphasis within educational institutions on the inclusion of this software in the undergraduate curriculum, especially in introductory circuit analysis courses for non-electrical engineering majors. For example, at ERAU, engineering students study the basics of MATLAB briefly as freshmen in the course "Introduction to Computing for Engineers". However, the use of MATLAB is not introduced in sophomore and junior years when students learn discipline-specific subjects. Although many senior projects involve heavy use of MATLAB, it is often difficult to refresh students' MATLAB skills at the senior level because they have not been consistently using the software.

In this paper, we describe our efforts in addressing this inadequacy in a junior level Electrical Engineering (EE) course by integrating MATLAB in the illustration of important concepts in basic circuit analysis. Our objectives are to enhance students' computer skills and improve their understanding of important circuit analysis concepts through graphic representations and visualization in MATLAB.

The effectiveness of this teaching practice is assessed through homework assignments and an end-of-course student survey. Collected assessment data confirmed the benefits of MATLAB usage in classroom instruction and its positive impacts on student motivation.

Our efforts were funded by ERAU Center for Teaching and Learning Excellence (CTLE) in Spring 2008 under "Advancing Teaching Practice Grant". It also serves as the initial study for ERAU College of Engineering's additional efforts to incorporate MATLAB component in undergraduate curriculum from freshman to senior years.

II. Implementation

Our teaching grant project was implemented in three sections of "Electrical Engineering I" in Spring 2008. This is an introductory course in circuit analysis and also the first course in Electrical Engineering (EE) taken by non-EE majors. The total number of students in these sections is 102. Majority are junior and senior students majoring in Aerospace

This work was supported by Center for Teaching and Learning Excellence at Embry-Riddle Aeronautical University, Daytona Beach, Florida.

T. Yang, I. Demirkiran, J. Liu, A. Helfrick, C. Grant are with College of Engineering, Embry-Riddle Aeronautical University, Daytona Beach, Florida 32114, USA (Corresponding author: Thomas Yang, phone: 386-226-7098; e-mail: yang482@erau.edu).

D. Pedersen is with Center for Teaching and Learning Excellence at Embry-Riddle Aeronautical University, Daytona Beach, Florida 32114.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_11, © Springer Science+Business Media B.V. 2010

engineering (AE), Mechanical Engineering (ME), Civil Engineering (CivE) and Engineering Physics (EP).

We used MATLAB to illustrate many important concepts in circuit analysis, including:

A. Solving linear systems of equations resulted from Nodal Analysis and Mesh Analysis

As fundamental circuit analysis methods, Nodal Analysis and Mesh Analysis utilize Kirchhoff's current and voltages laws to solve linear circuit problems. Often, it is tedious to solve the resulting linear systems of equations by hand. In MATLAB, the solution can be obtained via simple commands. In our project, we introduce two straightforward ways to perform this task, i.e., solution based on matrix inverse and solution based on Gaussian elimination.

B. Voltage and current across capacitors and inductors

The current and voltage waveforms are plotted and compared using MATLAB. The relationship between the waveforms, which involves differentiation and integration, is clearly shown when they are plotted in the same window. Therefore, the related formulas become easier to understand and memorize.

C. AC Circuit Analysis

The AC current and voltage waveforms are plotted in the same MATLAB window. The phase difference of the two waveforms, which is key to understand and analyze AC circuits, is easily seen.

D. First and Second order transient analysis for circuit behavior after switches change position

The model to describe transient process is differential equations. Merely solving the differential equations does not provide much insight into the dynamic evolution of the physical process. Using MATLAB, the transient responses are accurately plotted which give clear illustration of the physical meaning of the solution. Moreover, the forced and natural responses are also clearly shown together with the complete response.

E. Frequency characteristics of passive filters and the phenomenon of "Resonance"

Frequency response of passive filters is critically important to understand the principles behind modern electronic equipments. In our project, the magnitude response of the filter transfer function is plotted using MATLAB. Important concepts, such as resonance frequency, cutoff frequency, bandwidth, and quality factor, are all clearly illustrated in the plots. This makes the understanding of these concepts almost effortless.

Sample instruction materials are included in **Section III** of this paper.

After students were taught how to use MATLAB to plot desired graphs, they were assigned simple problems for practice purposes. Sample MATLAB codes were posted on Blackboard as examples. In MATLAB assignments, we emphasize that the results obtained from computer software should confirm, validate and match the results from analysis by hand. For example, the solution of a linear system of equations from Nodal Analysis obtained in MATLAB should match the solution obtained by hand. Also, in determining the types of passive filters, the answer can be found by comparing the magnitudes of the transfer function at DC (Direct Current) and at very high frequency. This answer should be confirmed by the magnitude response plot generated in MATLAB.

In final review of course materials, MATLAB was again used to assess students' understanding of key concepts. Here are some examples:

1. To evaluate students' understanding of the relationship between the voltage and current across capacitors/inductors, a multiple choice question is constructed with MATLAB. For example, a voltage waveform across a capacitor is plotted together with three possible current waveforms. Students are asked to identify which of the three is the current corresponding to the given voltage.

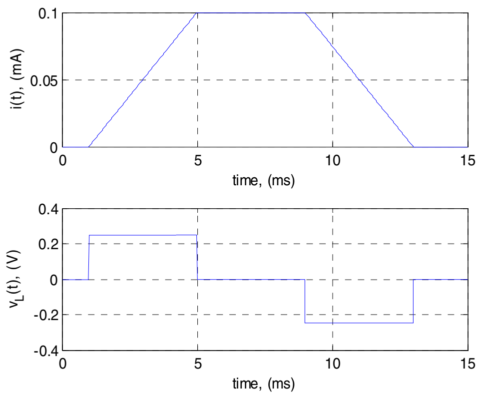
2. To evaluate if students are able to distinguish three types of second-order transient responses, i.e., overdamped, underdamped and critically damped responses, multiple plots are generated through MATLAB. Students are asked to determine the type of the transient response for each plot.

3. To assess students' ability to associate frequency selectivity of bandpass filters with quality factor and bandwidth, magnitude responses of multiple bandpass filters with same center frequency are plotted using MATLAB. Students are asked to order the quality factors and bandwidths of those bandpass filters.

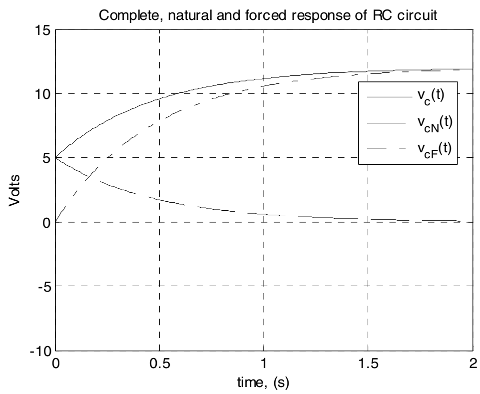
Sample review questions are included in Section 3 of this paper.

III. Sample Instructional Materials

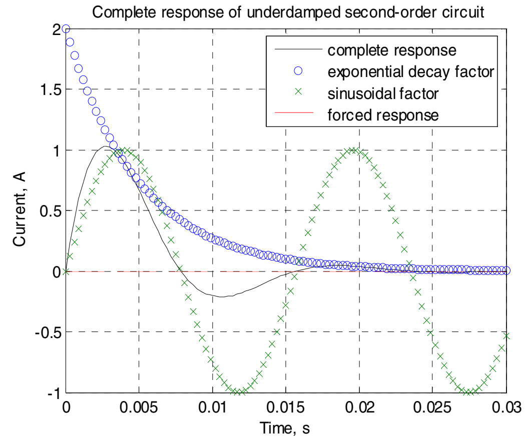
A. The relationship between an inductor voltage and current



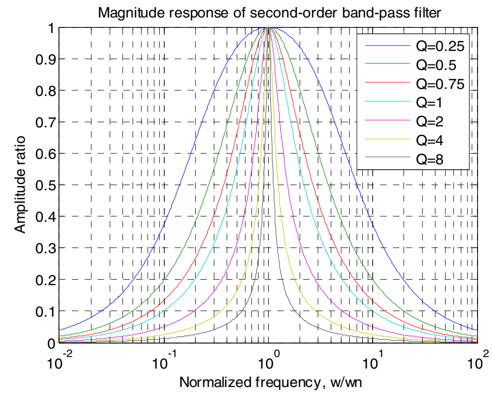
B. Complete, natural and forced responses of first order RC transient circuit



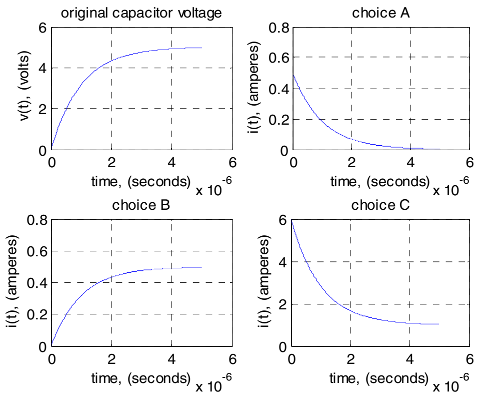
C. Response of a second-order underdamped transient circuit



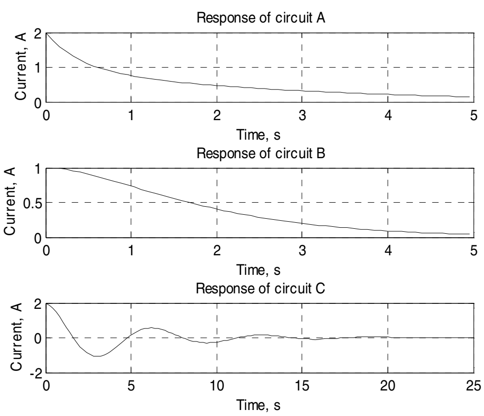
D. Using magnitude responses of second-order bandpass filters to illustrate the concepts of resonance, quality factor and bandwidth



E. Sample Review Question: given the capacitor voltage, choose the corresponding capacitor current



F. Sample Review Question: for each of the following figures, determine the types of second-order response (over-damped, under-damped or critically damped)



IV. Assessment and Outcome

Based on students' performance in MATLAB assignments, it was found that most students were able to perform basic mathematical operations in MATLAB to assist the solution of fundamental circuit problems. In addition, many students gained an understanding of the advantages and limitations of solution by hand analysis and solution via MATLAB. In particular, they realized that computer software is a powerful aid to problem solving, but by no means a substitution of theoretical analysis.

One of the major reasons we choose the course "Electrical Engineering I" to incorporate MATLAB illustration and visualization is that, unlike mechanical systems, electricity cannot be directly seen, heard or felt, especially for students who are new to the EE field. Thus, it is important for us to assess whether students feels more comfortable about this challenging subject under our new teaching practice. Assessing their subjective opinions is very helpful also because student motivation is of paramount importance in STEM (Science, Technology, Engineering, and Math) disciplines. Therefore, we designed an end-of-course survey to assess our teaching practice from the students' perspective. Below is the list of survey questions we used:

1. I can use MATLAB to perform matrix and vector operations.

2. I can use MATLAB to solve linear systems of equations resulted from Nodal Analysis and Mesh Analysis.

3. I can use MATLAB to plot data and mathematical functions such as voltage and current across resistor, capacitor or inductor.

4. I can use MATLAB to plot responses of first and second order transient circuits.

5. I can use MATLAB to plot frequency responses of passive filters based on transfer functions and determine the types of the filters.

6. I have a better understanding of relationships between voltages and currents across capacitors/inductors through visualization by MATLAB.

7. MATLAB plots of first-order transient responses help me understand the dynamic evolution of voltages/currents during transient process.

8. MATLAB plots of second-order transient responses help me understand the differences among underdamped, overdamped and critically damped transient processes.

9. MATLAB plots of passive filters' frequency responses help me understand the importance of transfer function.

10. MATLAB plots of passive filters' frequency responses help me understand the connection between filters' circuit structure and practical systems' user requirements.

11. MATLAB plots of various bandpass filters' frequency responses help me understand the concepts of quality factors, bandwidth, frequency selectivity and resonance.

12. Overall, I appreciate the incorporation of MATLAB into the instruction of Electrical Engineering I.

For each of the above questions, students' will provide a score ranging from 1 to 5, with 1 being "strongly disagree" and 5 being "strongly agree". Among these questions, 1-5 are concerned with students' ability to use MATLAB to assist circuit problem solving, and 6-11 evaluates the improvements in students' understanding of course materials. Question 12 is the overall evaluation of the teaching practice.

Survey results indicate that the majority of students believe the graphic illustration in MATLAB substantially improves their understanding of relevant course materials. The average score for question 12 was **4.5/5**. Specifically, the instruction of second-order transient analysis (question 8) and passive filters' frequency responses (questions 9-11) benefit most from visualization in MATLAB. Also, the plotting of voltage/current across capacitors and inductors (question 6) helps students quickly grasp the relationship between the voltage and current waveforms.

Also, we found students did appreciate the usefulness of MATLAB as a software tool to facilitate efficient solution of electrical engineering problems. However, concerning students' ability to perform analysis using MATLAB (questions 1-5), there is much variation among students' responses, with average scores varying between **3.5** to **4.2**. There are many students who are not proficient in using basic MATLAB commands. This further confirms the necessity of enhancing MATLAB instruction in engineering curriculum consistently throughout undergraduate years.

V. CONCLUSION/DISCUSSION

In this paper, we described our recent efforts in integrating MATLAB into the instruction of the course "Electrical Engineering I". It is concluded that the visualization of circuit analysis concepts in MATLAB facilitates instruction and promotes student learning. Meanwhile, the new teaching practice demonstrated to the students the advantages and limitations of computer software in engineering problem solving, and the relationship between the results obtained by software tools and by theoretical analysis.

There are some additional observations based on the findings of this project. Here we discuss two of them:

First, the basic MATLAB knowledge students learned in the freshman year still need to be substantially reviewed in our project, because students have not been using MATLAB consistently throughout the course of their college study. If more engineering courses, especially at the sophomore and junior level, incorporate MATLAB in their instruction, only a modest amount of effort is need to refresh students' MATLAB skills in each course.

Second, a balance needs to be achieved between the amount of course materials covered and the amount of time the instructor and students devote to MATLAB instruction and assignments. In promoting our teaching practice, we need to be cautioned against letting the MATLAB take too much of instruction time and student efforts in assignments. The course materials need to be the top priority, and the MATLAB element should not increase students' burden significantly. After all, student motivation should never be compromised, especially for STEM disciplines.

REFERENCES

[1] MathWorks, "Getting Started with MATLAB", Natick, MA, 1996.
[2] A. Haubold, "Matlab for First-year College Engineers", *The 37th ASEE/IEEE Frontiers in Education Conference*, pp. F1H-7 – F1H-12, 2007.
[3] MathWorks, "Learning Matlab", Version 6, Natick, MA, 2001.
[4] A. Azemi, C. Stoock, "Matlab for First-year College Engineers", *The 26th ASEE/IEEE Frontiers in Education* Conference, Vol. 2, pp. 599 – 602, 1996.
[5] M. Durán, S. Gallardo, S. Toral, R. Martínez-Torres, F. Barrero, "A learning methodology using Matlab/Simulink for undergraduate electrical engineering courses attending to learner satisfaction outcomes", *International Journal of Technology and Design Education*, Vol. 17, No. 1, pp. 55-73, 2007.
[6] A. Azemi, E. Yaz, "Utilizing Simulink and MATLAB in a graduate nonlinear systems analysis course", *The 26th Frontier in Education Conference*, Vol. 2, pp. 595 – 599, 1996.
[7] E. Yaz and Azemi, "Utilizing MATLAB in Two Graduate Electrical Engineering Courses", *The 25th Frontiers in Education Conference*, Vol. 1, pp. 2c6.1-2c6.4, 1995.
[8] S. Ayasun, C. Nwakpa, "Transformer Tests Using MATLAB/Simulink and Their Integration into Undergraduate Electric Machinery Courses", *2006 Wiley Periodicals Inc.*, pp. 142-150, 2006.
[9] V. Pires, J. Silva, "Teaching nonlinear modeling, simulation, and control of electronic power converters using Matlab/Simulink", *IEEE Transactions on Education*, Vol. 45, pp. 253-261, 2002.
[10] R. Patel, T. Bhatti, D. Kothari, "MATLAB/Simulink-based transient stability analysis of a multi-machine power system", *International Journal Electrical Engineering Education*, Vol. 39, pp. 320-336, 2003.

Istopolis: Design Issues and Evaluation' Results of an Integrated Web-Based Educational Application

C. T. Panagiotakopoulos
Departmnet of Primary Education,
University of Patras, Greece.
cpanag@upatras.gr

T. S. Papatheodorou
Departmnet of Computer
Engineering and Informatics,
University of Patras, Greece.
tsp@hpclab.ceid.upatras.gr

G. D. Styliaras
Departmnet of Cultural Heritage
Management and New Technologies,
University of Ioannina, Greece.
gstyl@uoi.gr

Abstract **– In this study the design issues of a web-based application, named Istopolis, are described. Istopolis is an integrated system aiming to support History and Culture courses of primary and secondary education in Greece. This hypermedia application system can be installed on a server allowing access both to students and teachers. Authorized teachers are able to modify or add modules with text, virtual material and video, and also modify or add scenarios towards an optimum exploitation of the initially specified educational modules. Furthermore, results of the summative evaluation obtained by a sample of active teachers are presented, showing that problems such as "lost in cyberspace" have been evaded and that, in general, educators are rather positive against the developed application in terms of usability, quality of media and aesthetics.**

I. Introduction

Online systems aiming to support education are firstly met in the decade of '60 with the appearance of the first online bibliographic retrieval system [1]. Systems developed from 1963 to 1995 could be classified as the first generation of educational online systems.

After the wide popularization of Internet (April 1995), online educational systems met an increasing implementation, and can be classified in two categories [2]: (a) oriented to teacher-to-student communication [3] and (b) oriented to student-to-student communication [4]. The generic attributes of systems from both categories present many similarities with those of the first generation online educational systems.

Along with the emergence of the second generation of systems an effort to support the learning process took place. This effort was mainly based on elements of the traditional class [2].

With the wide diffusion of Internet and WWW, online systems started to be employed as sources of information retrieval [5] and means of design and development of learning environments [6] [7].

The third generation of web-based learning systems or online educational systems, as they are met in the literature, is based on novelties regarding the means of control and the direction of information flow [8]. Contemporary educational online systems are nowadays exclusively based on Internet. Along with these systems, the term "virtual teaching" emerged [9].

These systems offer exceptional and continuously enhanced capabilities to educators, who eventually have the potential to communicate with their students and assign them lessons by using educational scenarios as well as exercises and projects. Exercises and projects can be reviewed online, providing students with feedback in terms of educational comments and grades [2] [10]. Thus, the actual virtual class can be realized.

In the present work, the structuring (design and development) of the Istopolis application is presented. This application constitutes an integrated web-based hypermedia educational system aiming to support History and Culture courses of the Greek primary and secondary education. The structuring and interconnection of the various sub-units of Istopolis are also described, which aim at serving all the users (e.g. students, educators and administrator) interacting with the system. Finally, results of a predictive evaluation are presented, in terms of usability, aesthetics, quality of included media, and hierarchy of learning effect of differentiated information transmission media. The data used to perform this evaluation, were acquired from a sample of 36 educators, who used the developed system in real conditions.

II. Istopolis Application: Identification

The name of the presented application comes from parts of the Greek words Isto-ria (history) and Poli-tismos (culture). It is an integrated hypermedia web-based educational system constructed during 2007-2008 in the context of the Unit "Chrysallides" (Development of integrated educational activity products), of the Action "Pleiades" (Development of Educational Software and Integrated Educational Activity packages for Greek Primary and Secondary Schools & Distribution of Educational Software Products to Schools), with the Research Academic Computer Technology Institute (RACTI http://www.cti.gr) as the organization responsible for implementation monitoring. It was funded by the Greek Ministry of Education & Religious Affairs and the Operational Program for the Information Society, Measure 1.2, 3rd Community Support Framework.

The system was integrated through a continuous formative evaluation from the respective organization responsible for monitoring and controlling (i.e. RACTI) technical and educational issues. It was finally positively evaluated by the Greek Pedagogical Institute from a technical as well as an

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_12, © Springer Science+Business Media B.V. 2010

educational perspective.

The developed application can be installed on a networked computer, which educators and students are both able to access. It is worth noting that Istopolis is a content independent application.

The application's integration was realized through the cooperation of several specialists from various fields, such as educational design, programming, video creation, sound, video and image processing, text editing, design editing, educational scenarios' construction, content writing and text narration.

The generic internal structure of the system is illustrated in Fig. 1.

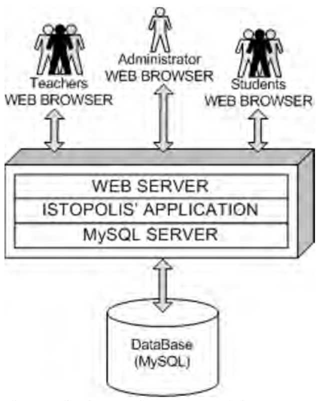


Fig. 1. The internal structure of the system.

The administrator can assign use rights to educators and register their properties such as the school and the class they are teaching to. Each educator in his turn can assign use rights to students of the classes he is teaching to. Through the Istopolis environment, educators are also able to assign to the students of a class, one or more scenarios for processing purposes, in order to actively engage them in the application's content.

The software is accompanied by a Content Management System (CMS), which addresses the need for:

a) Current content improvement

b) Inserting new integrated modules

c) Inserting new scenarios towards a more effective content exploitation.

The administrator is capable of assigning specific use rights to appropriate educators who will have the responsibility to operate the CMS.

A. Content

The student may have access to the content that is comprised of:

(a) 78 modules containing data regarding the Ancient, the Byzantine and the contemporary History and Culture. For the creation of these modules the following were used:

- approximately 400 pages A4 size, of original text written by archaeologists, historians and philologists,
- approximately 4000 pictures with legends and references in their original source,
- approximately 400 min of video, distributed in video clips lasting from 3 to 6 minutes,
- links to webpages with extra visual and text educational material relevant with the content of each module.

(b) 40 two-houred educational scenarios for the software exploitation in a school class level, which also determine the level of use, that is the educational grade that this software will be used (primary education – secondary education).

B. The educational scenarios

Each authorized student can easily explore and study the content of the application. The effective exploitation of the content though, is realized with the assistance of the originally created educational scenarios that come along with the software or by subsequent scenarios manually created from authorized educators. Creation of new scenarios is always necessary in order to assist the learning process of specific students or specific classes with alternative methodologies, addressing their personal skills, preferences and needs.

The scenarios essentially comprise of activities that should be performed by students of a class and include exercises, written projects (assignments) as well as instructions for their effective fulfillment. This way, the student becomes familiar with the educational content of Istopolis.

The scenario's instructions encourage the students to go back to specific software units or external links, study text documents or original sources, in order to be able to work out the exercises or compose the projects they were assigned.

The exercises included in the scenarios can be classified into three types: open, multiple choice and one-to-one correspondence. The application has the potential to automatically evaluate the multiple choice exercises as well as the one-to-one correspondence ones, depending on the settings designated by the scenarios' composer in advance. In this direction, the teacher may change the weight of each question or exercise, before they have been assigned to a class, determining an alternative weighting list.

Assignments are composed by the student through a proper environment facilitating a small hypertext editor that incorporates all the common capabilities for text processing. The small editor also supports interconnection of text with other kind of information (hyperlinks), located in the system's content or in the World Wide Web. Finally, it provides capabilities of image insertion and integration with the under-process-text.

As the above show, the application-user interaction is boosted in a considerable level.

III. IMPLEMENTATION DETAILS

The system was built by using HTML, PHP, Adobe Flash, Java, Javascript and MySQL. Apart from Flash, all software employed is based on open and acceptable standards [11] and runs in a school laboratory environment. A MySQL database stores the system's content along with user permissions and parameters about assigned projects. The system's content is generated on the fly by using PHP scripts that draw the appropriate content from the database. Exercises are implemented with Javascript and Java. The timeline is based on Javascript and Dynamic HTML, whereas Flash is employed for implementing the map interface. Flash communicates directly with the database via Javascript in order to display the appropriate map and the points on the map. Flash is used also for displaying the virtual interface and the menu on the upper right of every screen. Finally, the content management interface is based on PHP scripts.

IV. STUDENT'S INTERFACE

The application's information flow, from the user's perspective, depends on the kind of the user (administrator, teacher, and student).

The general pattern of the information flow concerning the student's access to the software's content is shown in Fig. 2.

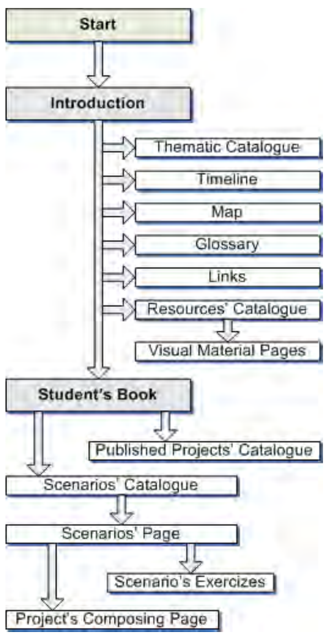


Fig. 2. The data flow of student's interface.

A. Content's access

The student can access the software's content in three different ways – choices. These choices are the "Thematic Catalogue", the "Timeline" and the "Map" (Fig. 2, Fig. 3).

The shortest way to access the content is through the "Thematic Catalogue" choice. Within the "Thematic Catalogue" modules and complementary texts are classified through their positioning in a thematic hierarchy.

With the "Timeline" choice, modules and complementary texts are classified through their positioning in chronological periods. This classification enables both unit and text search within a specific period of time, determined by the user.

Finally, through the "Map" choice modules and complementary texts are classified according to their association with a geographical area. If the module or the text can be associated with a certain location, it then appears positioned as a point on the map. Otherwise, it appears under the map of the region it was associated with.

On the other hand, the user can access a specific content in a combinative way through the "Timeline" and the "Map". In other words the user can access a specific content through the combined use of time and space.

As many researchers in the domain of hypermedia applications have already reported, due to several design issues, the phenomenon of "lost in cyberspace" has been observed. This phenomenon disorganizes and misleads the user [12]. To avoid such problems, we did not connect with hyperlinks any words or phrases within the main and the complementary text of each module. Instead, we used a term dictionary in every module that the user can access by pressing a button in each module's environment. All the small dictionaries of the modules act as subsets of a bigger dictionary that the user can find in the root menu and approach it alphabetically. Moreover, all hyperlinks of the module's menu operate and expand inside the specific module, apart from the "Resources", the "References" and the "Links".

Fig. 3. An actual screenshot of the student's interface.

B. Student's Book

An important "field" of the application from the student's perspective, is the "Student's Book". In that "field" the student can find:

a) The scenarios that his teacher sent him for processing.

b) The tasks that the student has fulfilled (exercises and assignments) and wants to be published in order to be evaluated by his teacher.

Note, that the field "Student's Book" is accessible by all authorized users. This fact emerged under the consideration that in this way, a better competition between the students will be created and maintained since all of their projects and worked out exercises will be in public view.

Moreover, in that field every student can view the evaluation of the projects submitted. Evaluation includes both the project grade and the comments justifying the received grade.

C. The form of the modules

All units' content appears in a specific structure, which allows the student to have easy access (always with less than three mouse clicks [13]) in every sub section, which either is in text (written text, scanned prototype text) or visual form (pictures - videos).

Every module consists of a main title and text, a video presentation, a wide group of high definition images and maps, original sources (if such exist), bibliography, complementary texts, interconnections with other module's content or with World Wide Web's URLs and a small dictionary hosting the module's terms.

Graphically, each module's content structure appears in Fig. 4.

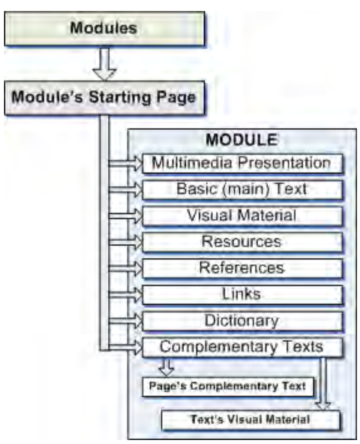


Fig. 4. The data flow structure of every module.

V. TEACHER'S INTERFACE

The teacher can be informed by the system for the scenarios already assigned to a class in order to be effectuated along with those that remain unassigned. He is also able to extract a scenario in XML format, change or enhance it and import the modified scenario back to the server. The online tools the teacher is offered are associated with:

a) The class administration, such as: assignment of a scenario or a set of scenarios to a class, establishment of an appropriate hierarchy for the scenarios to be accomplished when more than one scenarios are assigned to a class, collection of the student's projects, evaluation, and provision of grades and constructive feedback to the students.

b) The construction of new scenarios.

Information flow of the teacher's interface is depicted in Fig. 5:

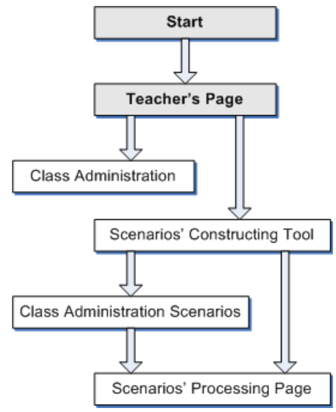


Fig. 5. The structure of data flow in teacher -mode.

The interface through which the teacher is able to perform the above tasks is illustrated in Fig.6:

Fig. 6. An actual screenshot of the teacher's interface.

VI. INTERACTION

According to several researchers [14], the interaction between educational software and user is an essential element positively contributing to the learning process. Istopolis offers several capabilities concerning the interaction with the student, which include:

1. Navigation in 3D space in order to discover the software's research and configuration tools, after the student has logged in. This way the students are welcomed by an environment which follows a gaming fashion that is more close to the student.

2. Content search using simple research methods, such as the modules' catalogue.
3. Content search using complex search tools, coupling intense semantic charge with simple graphic interface (such as the Thematic Catalogue, the Timeline or the Map) and exploit the intrinsic hierarchical information structuring.
4. Content search through complex combinations of search criteria regarding "subject - location - time".
5. Access to same content through several paths.
6. Project writing through a small hypertext processor with complete editing capabilities and project publishing.
7. Content collection, temporary storage and reuse in the projects.
8. Fill out and submission of self-evaluation exercises.

The above are only the most important from the whole of the interactive functionalities offered to the students.

Regarding the student-educator interaction, it is noted that the software is designed to be used inside the class, while the educator is present supervising the students.

VII. Application's Evaluation

The application, during its implementation, was structured through a continuous formative evaluation by the supervising organization as well as the development team. The summative evaluation was performed with multiple groups of users and multiple evaluation subjects.

In the following, the results of a predictive summative evaluation received from a sample of 36 active teachers of primary and secondary education are presented.

The evaluation procedure was performed with the sample divided (for practical reasons) into two groups. The software was initially presented to each group by one researcher for two teaching hours, while each participant was working at a personal computer. After the presentation the participants were asked to work with the developed software for three teaching hours. Finally, each participant was asked to provide feedback on specific questions filling an appropriate anonymous questionnaire.

The variables of the scale used for the evaluation [14] [15], aimed in collecting data regarding the applications' usability, the aesthetics, the quality of the media included, and the establishment of an hierarchy for the learning effect of the information transmission media from the teacher's perspective.

Analysis of the self-evaluation of the participants' knowledge in basic issues of computer use (Microsoft Windows/Office, Internet) showed an average of 3.9 in the five grade Likert scale (knowledge: none = 1, 2, 3, 4, 5 = excellent) with st.d. 0.7.

A. Usability

We asked the sample to provide answers to specific questions with expected answers within the five-grade Likert scale (not at all = 1, 2, 3, 4, 5 = absolutely or difficult = 1, 2, 3, 4, 5 = easy). The questions, the answer's average, the standard deviation and the mode (the value that occurs with the highest frequency) with the respective rates are shown in Table 1.

TABLE 1
QUESTION, MEAN, STANDARD DEVIATION AND MODE

Question (usability)	Mean	St.D.	Mode
U1. At the very beginning working with Istopolis was:	3.97	0.97	5 (36.1%)
U2. In general, you found the software's handling:	4.13	0.76	4 (41.7%)
U3. Did you understood the way this software operates:	3.92	0.81	4 (52.8%)
U4.Do you think that the students will be able to use it:	3.58	0.69	3 (52.8%)
U5. Can you create a scenario with exercises and projects through the software:	3.56	0.98	4 (33.3%)
U6. Can you assign a scenario to the students:	3.92	0,84	4 (47.2%)
U7. The software favors interaction between student and content:	4.17	0.74	4 (52.8%)
U8. Navigation in the software's units was performed:	4.44	0.59	5 (50.0%)
U9. The use of the search tools (Thematic Catalogue, Timeline, Map) is:	4.21	0.72	5 (69.4%)
U10. Functionalities of the "Student's Book" for the teacher and the student is in terms of usability:	4.03	0.61	4 (63.9%)
U11. Did you feel "lost" at any time, while navigating in the software:	1.50	0.56	1 (52.8%)
U12. Did you know at all times your exact position in the software:	4.19	0.78	5 (41.7%)

Usability effects were analysed using a chi-square test. Teachers found the software easy to handle (U3: x^2=17.11; df=3; $p< .001$). This result is replicated when teachers comment whether their students will be as well able to handle it (U4: x^2=9.5; df=2; $p< .01$). As a result, the high level of usability makes educators feel confident to assign scenarios to their students (U6: x^2=12.67; df=3; $p< .01$). From the teachers' point of view, the navigation through modules poses no usability constraints (U8: x^2= 17.11; df=3; $p< .001$).

In addition, the sample was asked to answer two more questions:

- In general, were there any points, where due to difficulties in using the software you finally did not work with it?

- In general, did you find any problems using the software or not?

Answers regarding the 1st question were 97,2% negative. This means that only one subject claimed that there was a point where working with the software was not possible due to difficulties in its use. However, when this subject was asked with a follow-up question to spot this point, no answer was given. Answers regarding the 2nd question were 100% negative.

The findings described above, show that the participants in the evaluation believe that the software is easy and flexible in its use. Furthermore, the software's design met with success regarding the aim to avoid the "lost in cyberspace" phenomenon.

B. Aesthetics and quality of media

The sample was asked to evaluate the software's aesthetics and the quality of information transmission media used by it. The outcome is presented below (Table 2):

TABLE 2
QUESTION, MEAN, STANDARD DEVIATION AND MODE

Question (aesthetics – quality of media)	Mean	St.D.	Mode
A1Have you found image and video quality fair enough?	4.08	0.97	5 (41.7%)
A2. In which level did you find the software's images good regarding their quality?	4.02	0.91	5 (38.9%)
A3. Was the display's color aesthetics appealing?	3.92	0.73	4 (47.2%)
A4. How much distinctive – readable was the text?	3.97	0.81	4 (36.1%)
A5. Are the design and the presentation of the application's screen appealing?	3.94	0.83	4 (41.7%)
A6. Is there cohesion regarding the presentation of the screen in every module?	3.89	0.75	4 (52.8%)
A7. In general, how do you consider the software from an aesthetics perspective ?	3.94	0.86	4 (36.9%)

In general, application's aesthetics were assessed as being of high quality (A7: x^2=9.78; df=3; p< .05).

Establishment of a hierarchy regarding the quality of the information transmission media' factors (text, pictures, graphics-maps, videos) was analyzed with the assistance of the Friedman's test and showed an important advantage of video (x^2=25.43; df=3; p< .001) followed by pictures (Table 3).

C. Hierarchy of information transmission media' learning effect

The sample was asked to give their opinion for which information transmission media is considered as helpful for the student in order to better understand the concepts in the developed application. Data analysis through the Friedman's test showed again an important advantage of video (x^2=19.23; df=3; p< .001) followed by pictures (Table 3).

TABLE 3
RANK OF MEDIA' QUALITY AND EFFECT

Media	Quality - Mean Rank	Effect - Mean Rank
Videos	1.92	1.72
Pictures	2.06	2.53
Graphics – maps	2.78	2.83
Text	3.25	2.92

The previous results show for one more time that an image on its own is able to paint a thousand words. For this reason, during the design of an educational material this should be seriously taken under consideration [16].

Finally, all participated teachers, who answered a relative question, stated that they would like to use the developed software in their classroom, as part of their everyday teaching practice.

VIII. CONCLUSIONS

In this study, issues regarding the design and development of the application named Istopolis were analyzed. This application, in its current form, comprises an integrated platform aiming to support History and Culture courses of primary and secondary education in Greece. It is a web-based application offering the potential to modify or enhance its content through a CMS. Using the presented application, the educators are able to assign exercises and projects to their students through educational scenarios, evaluate them online and provide them with grades and constructive feedback.

The results of the final predictive summative evaluation with a sample of 36 teachers showed that the developed application benefits the educational process while being handy, very appealing in terms of aesthetics and of high quality with regards to the quality of information transmission media.

REFERENCES

[1] C. P. Bourne, and B. H. Trudi, "*A History of Online Information Services, 1963–1976*", Cambridge, Mass.: MIT Press, 2003.

[2] R. Sheizaf, B. Miri, D-G. Yuval, and T. Eran, "QSIA – a Web-based environment for learning, assessing and knowledge sharing in communities", *Computers & Education 43*, (3) 273–289, 2003.

[3] Y. J. Dori, M. Barak, & N. Adir, "A Web-based chemistry course as a means to foster freshmen learning", *Journal of Chemical Education, 80*(9), 1084–1092, 2003.

[4] M. Guzdial, and J. Turns, "Effective discussion through a computer-mediated anchored forum", *The Journal of the Learning Sciences, 9*(4), 437–469, 2000.

[5] C. A. Merlic, B. C. Fam, M. M. Miller, "WebSpectra: Online NMR and IR Spectra for Students", *Journal of Chemical Education, 78*, 118–120, 2001.

[6] Dori, Y. J.; Barak, M., "Virtual and physical molecular modeling: Fostering model perception and spatial understanding", *Educational Technology & Society*, 4, 61–74, 2001.

[7] Donovan, W. J.; Nakhleh, M. B., "Students' Use of Web-Based Tutorial Materials and Their Understanding of Chemistry Concepts", *Journal of Chemical Education, 78*, 975–980, 2001.

[8] Nachmias, R., Mioduser, D., Oren, A., & Ram, J., "Web-supported emergent-collaboration in higher education courses", *Educational Technology & Society*, 3(3), 94–104, 2000.

[9] Sheremetov, L., & Arenas, A. G., "EVA: An interactive Web-based collaborative learning environment. *Computers & Education, 39*, 161–182, 2002.

[10] Panagiotakopoulos, C., Papaterpos, C., & Papatheodorou, Th., "Presentation of a hypermedia system supporting History: Basic characteristics and content's evaluation", *Proceeding of 1st Educational Conference for the Introduction and Use of ICT in Education* (P. Politis, ed.), 759-765, 2009.

[11] Rafaeli, S., Barak, M., Dan-Gur, Y. & Toch E., "QSIA – a Web-based environment for learning, assessing and knowledge sharing in communities", *Computers & Education, 43*, 273–289, 2004.

[12] Otter, M. & Johnson, H., "Lost in hyperspace: metrics and mental models", *Interacting with Computers, 13*(1), 1-40, 2000.

[13] Panagiotakopoulos, C., Pierrakeas, C. & Pintelas, P., "*Educational Software Design*", Patra: Hellenic Open University Publications, 2005.

[14] Picciano, A.G., "Beyond student perceptions: Issues of interaction, presence, and performance in an online course" Journal of Asynchronous learning networks, 6(1), 21-40, 2002.

[15] Panagiotakopoulos, C., Pierrakeas, C. & Pintelas, P., "*Educational Software and its Evaluation*", Athens: Metaihmio Publications, 2003.

[16] Spiliotopoulou-Papantoniou, V., Karatrantou, A., Panagiotakopoulos, C. & Koustourakis, G., "Visual representations of the internet in greek school textbooks and students' experiences", *Education and Information Technology, 14*, 205-227, 2009.

TEAC^{2}H-RI: Educational Robotic Platform for Improving Teaching-Learning Processes of Technology in Developing Countries

Juan J. González[1], PhD Jovani A. Jiménez[2], PhD Demetrio A. Ovalle[3]
National University of Colombia
Carrera 80 numero 65-223 (M8A-309) Medellín-Colombia
[1] jjgonzaleze@unal.edu.co, [2]jajimen1@unal.edu.co; [3]dovalle@unal.edu.co

Abstract—**The technological development achieved in each of the epochs of the human history has changed the means of production, communication, transportation and education. There are many results because of the incorporation of technology on education, including remote laboratories, online courses, and educational robotics. In recent years, the latter has become very popular because of its significant contributions to the learning process, strengthening of cognitive skills, including creativity, design and problem solving skills. However, this phenomenon has been focused in the developed countries, and many of the applications of the developing countries rely on a technological transference from the developed countries. This model is not the most suitable due to the fact that the tool should be adapted to the environment and its user, and not the opposite. This article presents some of the main contributions in this issue and outlines the TEAC^{2}H-RI kit. This kit was developed based on the Colombian educational environment characteristics and the experience in robotics of the research group GIDIA. It will be used in Senior High School Students.**

***Index Terms*—Society and Computing, Educational Robotics, Robotic Swarm, Constructivism, Construccionism**

I. Introduction

As well as the societies evolve so do its means of production, communication, transportation and education. In the last one, one of the most important moments was in the XIX century when many researchers proposed the switch from the passive paradigm to an active paradigm where the participation of the learner is more active. It comes as a result of the works of pedagogues, educators and psychologist, including, Johann Heinrich Pestalozzi, Friedrich Froebel and Maria Montessori [1]. These previously mentioned approaches, along with the subsequent, are the root of the development of the constructivism theory of the Swiss Psychology Jean Piaget. In this theory the knowledge is not assume like a fluid which can be transferred from one person to another, but like a process where the knowledge is created in the learner's mind [2]. Based on it, the South African mathematician Seymour Papert proposed the Construccionism Pedagogy, which also claim, as the constructivism theory, that the knowledge is not transferred but created in the learner's mind, but additionally, that the method to achieve that is through the construction of something tangible by the learner [3] [4].

Papert and his research group at the Massachusetts Institute of Technology (MIT), based on the Construccionism Pedagogy, developed a programming language for kids, called Logo. Logo had a Graphic User Interface which allowed programming a turtle robot. Afterward, LEGO and Fischertecnnik with a similar concept developed kits to build robots using many bricks, including tires, gears, lights, sensors. The Media Lab of the MIT and other institutions used these kits to explore the results of the interaction with them. Most of the researchers concluded that theses kits not only are useful for the learning process but also for the strengthening of the creativity, design, troubleshooting, team work and social awareness, among other cognitive skills.

These kits produced by Develop Countries have been implemented in these regions, and in some cases in Developing Countries. Although, the last one have social-economic conditions very different from those of the formers, then the performance of these kits without the correct reading of the context could generate no satisfactory results. It is because by this methodology the user and his environment has to be adapted to the educational tool, but it is the tool that should be adapted to the user and his environment. This phenomenon is very common in the technological implementation in the Latin American Countries and it is very criticized by many different authors. [5]

In order to provide an autochthonous solution to this problematic the Artificial Intelligence in Education Research Group of the National University of Colombia addresses these difficulties in the research project: Educational Robotics: Intelligent Machine in Education.

This paper is organized as follows: Section II briefly reviews some approaches in Educational Robots. Section III presents some cases of the implementation of educational robots in developing countries. Section IV describes the

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_13, © Springer Science+Business Media B.V. 2010

proposed methodology. Section V presents the kit TEAC2H-RI (Technology, Engineering And sCience eduCation through Human-Robot Interaction). Finally, section five presents the conclusions and future work.

II. Educational Robotics

Since the creation of the Logo Turtle in the 60's by the Media Lab MIT until our days the advancement in Educational Robotics has covered not only the primary school students but also the high school, undergraduate and graduate students, and even the industry workers. Some of the most important educational kits are: LEGO Minstorms, Logo, Fischertechnik [6], Pioneer, K-Team, IroRobot y Handy Board [7], [8], [9]. The three formers are used in education and toy industry. On the other hand, the last five could be implemented in the school, university and in the industry. These kits are very useful to explain and verify algorithms, to learn some concepts of programming languages and to encourage the strengthening of different cognitive skills [10].

From the different options available LEGO Minstorms and Fischertechnik are the most widely used in the world because its relative low cost, in comparison with the others, and the educational resources that provide. The philosophy of both is very similar because is based on interlocking bricks. One of these kits is RoboExplorer. It is manufactured by Fischertecnik, and it is consists of ultrasonic range sensors, a color sensor, a track sensor, a NTC resistor and a photoresistor. The locomotion is based on two servomotors and a crawler drive. Also, with this kit is provided the ROBO interface, and the programming software ROBOPro [11]

As can be seen the characteristics of these kits made of them a very useful tool in the teaching of important subjects as mathematics, physics, electronics, and algorithms. These kits at the university, in careers related with Computer Science, contribute with a deeper understanding of the algorithms concepts. On the other hand, in industry there are many applications, including the process simulation, training in the control machine. However, its implementation must be adjusted to the particular context of the region. The following section presents some approaches in this matter.

III. Educational Robotics in Developing Countries

The great deal of problems in the majority of the developing countries, especially in the Latin American Countries, requires that the inhabitants face them not only with a solid knowledge in their speciality area but also with high research skills. Furthermore, it is necessary to increase the attraction of school students for careers related with technologic production, such as engineering. Statistically (see Table 1), it can be seen that there is a research crisis in Colombia and in almost all the Latin American Countries, e.g. in Colombia in 2006 there was only 151 researchers per million inhabitants, whilst for the same year in USA and Japan there is 4663 y 5573 researchers per million inhabitants, respectively [12]. In the table I, can be seen the number of researchers per million inhabitants in developed and Latin American countries. There is very clear that the research crisis covered almost the whole region, and is less severe in Chile, Brazil, Argentina and México. The research capacity of any country is one of the most important resources to affront the social, ecological, technological, and economical challenges. Therefore, it can be understood why the countries with a low number of researchers per million inhabitants, have a lower progress than those with a greater number of researchers.

TABLE 1
Researchers per 1 000 000 Inhabitants (Taken From [12])

Researchers per 1 000 000 inhabitants (FTE)	
Argentina	980
Bolivia	120
Brazil	629
Canada	4157
Chile	833
Colombia	151
Ecuador	69
Guatemala	25
Japan	5573
Mexico	460
Panama	62
UK	2881
USA	4663

Aware of this fact, some research centres and universities have implemented different educational robotics courses, to find a solution to these issues. Some of them are the Carnegie Mellon University (CMU) and the MIT. Next is presented some of their results

In Qatar and Ghana the Carnegie Mellon University associated with some higher educational institutions of these regions, developed a course on autonomous robots with the aim of strengthening the knowledge on algorithms and the creativity in the undergraduate Computer Science students. The educational robotic tool was the MIT Handy Board, the software was the Interactive C and the hardware the bricks of LEGO. At the end of the courses was concluded that robotics as an educational tool in Computer Science meet the proposed objectives and also increase the interest of the learners by technological topics and research. Another important conclusion was that the interaction, for the goal accomplishment, among the team members strongly depends on the cultural environment. This fact has to be observed at the design stage of a course in this thematic [13].

In Costa Rica the Massachussets Institute of Technology (MIT) worked with a school located in one of the rural zone of this country. The educational robotic tool was the LEGO Mindstorms and the software was the Yellow Brick Logo. From the results some of the conclusions were that educational robotics is a very useful tool for the student learning process support and also for the strengthening of the creativity and the troubleshooting skills. [14].

As was illustrated before there are many robotic educational tools, however, because they were developed in countries with socio-economical conditions very different from the Latin American countries, its effectiveness is lower in those countries than a autochthonous propose made from a fair reading of the context. Moreover, they are very expensive in comparison with the economical resources of the majority of the educational institutions of these countries, restricting the widely spread of this tool. Due to this fact, it is necessary to make a deeper reading of the Latin American reality and bring an autochthonous solution according to the region necessities [5]. From the educational view, one important constraint of the majority of these kits is the specific number of pieces that they have, which influence the mind of the learner to develop some specific prototypes. Technically, one of the most important constraints is that each brick has a group of equations that rules its behavior which can not be modified by the user [15], additionally the learner can not integrate new bricks of his/her own creation. These constraints restrict the learning process, and the strengthening of creativity and design.

IV. Proposed Methodology

The Artificial Intelligence in Education Research Group of the National University of Colombia, addressed the difficulties mentioned above through the research project, *Educational Robotics: Intelligent Machine in Education*. This project is ongoing. In this project, based on a objective view of the Colombian reality and the experience in robotics of the Research and Development Group in Artificial Intelligence of the same institution, was developed the robotic swarm TEAC^{2}H-RI (Technology, Engineering And sCience eduCation through Human-Robot Interaction) which will be implemented in the Senior High School students of the Metropolitan Area of Medellin.

In this zone will be chose ten schools and to each of them will be delivered a kit TEAC^{2}H-RI, composed of 4 robots. To be eligible, each school must fulfil with some minimum requirements, i.e. a computer room. If the school not fulfils this requirement, our university will provide a special room for this purpose. Each course will have duration of 20 hours distributed in five weeks, with seven workshops.

The aim of the first workshop is to get the commitment of the teachers with this process, by means of the increasing of their comfort with the topics of the course and dismissing the negatives or technophobic paradigms. This will ensure that, even when the course has finished, the tool remains its pertinence to be used in the institution.

For this purpose, it will be presented in the workshop, the topics to be studied and the expected results. Additionally, it will be illustrated some pedagogies and theories of active education, such as the Piaget's Constructivism theory, the Ausubel's learning theory and the Contructionism of Papert. Afterward, it will be exposed and argued the reasons to use active pedagogies of education as their teaching methods supporting this in the social, economical and productive necessities of the Colombian context. Finally, it will be demonstrated that the active role of the student in his/her learning process can be done through the interaction with the kit TEAC^{2}H-RI.

In the second workshop the objective is to increase the motivation of the learners for robotics, especially for the kit TEAC^{2}H-RI. The success of the course depends on this fact then it should be handled with caution. First of all, we will present a movie or documental about robotics. Then, they will be induced to observe and analyze the behavior of the robotic swarm. After, teams of four students will be formed and each of them will work with a tutor and a robot. It is done in order to make of robotics meaningful material for them [16].

In the third workshop the objective is that the learners understand the theories that rule the behaviour of the robots through the modification of the parameters of them. The tutor will stimulate the interest of the learners to interact with the kit, and also he/she will provoke situations where the learners could conclude some important theories that rule robotics. For example, the tutor will incite the learners to move the batteries position and observe how the change in the centre of mass affects the robot's stability. There are many other important concepts that could be understood through the interaction with the kit, including overlapping of light beams, angular momentum, friction, resistance [17]. Some of the skills that will be strengthened are analyses, problem solving, and the team work.

In the next workshop the objective is that the teams design, and develop an addition, improvement, or a modification on the robotic platform. Also, the learners could choose to design and develop a small automation system which must be associated with the topics of the course. For both activities, the tutor will be a supporter to the learner during his design and construction stage. The skills to be strengthened in this workshop are the design, cooperation, creativity and research.

In the fifth workshop the objective is to strengthen the writing skills. In order to do that, each student will be asked to write an article, which must contain abstract, introduction, theoretical support and their experience in the previous workshops. Other issues that have to be mentioned in the article are: "what are the characteristics of the device that you developed in the previous workshop?", "Name an industrial sector where it can be implemented", "Why did you choose this design over others?" among others. Finally, the article should have the conclusions and references section. This strategy is used because people like to write about their interests [18].

The last objective is to promote the research discussion spaces among the students. For this purpose, each group will present, in front of their partners, the results obtained in the fourth workshop. The other students in the audience will be encouraged to criticize constructively the projects, to let the team to identify the improvements or limitations of the prototype. It is also with the aim to strengthen the team ability to appreciate the diverse points of view, and as a consequence expand his community sense [17], [13].

Finally, the Artificial Intelligence in Education Research Group, based on the previous experience in educational topics, will develop the test to evaluate the effect of the course in the learners, testing the performance of the students before and after of their participation in the course.

Despite de fact that the course finishes in this stage, the support of our group will continue, because through the web page, http://xue.unalmed.edu.co/~dbetanc0/index-3.html, we will present updated information about the kits, such as designs, algorithms or manuals that could improve the performance of the robot. Furthermore, the students could interact with the Developer research group of the kit through the web page or the e-mail, facilitating in this way the feedback of the concepts and validity of the kit through time. Next, it is mentioned some software and hardware characteristics of the kit TEAC^{2}H-RI.

V. Educational Robotic Kit: TEAC2H-RI

TEAC^{2}H-RI is a robotic swarm composed of four robotic agents, i.e. three sons and one mother. It is a hybrid system because it has distributed and centralized characteristics. It is distributed because the son-robots, with similar behaviour as the mother-robot, could interact among them. It is centralized because the mother-robot could perform the role of group leader. See Fig. 1 where a mother-robot is interacting with a son-robot.

The robots not only interact among them, but also with the user by means of sound signals (buzzer) and light (LEDs). It facilitates to the user to conclude if the sensor connections or the robot behaviour is appropriated, and hence take the corrective actions [19].

Each robot, when has been properly configured by the user, it could carry out navigation activities and obstacle avoidance. Even so, the mother-robot is the only one that could also send and receive the information from and to a remote PC, either by way of a radio link or a data cable. The modes of navigation are: track follower, by means of the QRB1114, obstacles avoidance through infrared sensors or touch sensors.

Fig. 1 Mother and Son Robot Interacting

This robotic swarm will be used in the workshops with the schools. In each workshop the maximum number of participants is 16, divided uniformly in four groups and a robot assigned for each team. This work topology with the kit TEAC^{2}H-RI facilitates not only the cooperation among the members of each team, but also the cooperation among the groups, making with this that the group and intergroup interaction skills be strengthened. This is the fundamental model of relations in enterprise and research projects. One deeper explanation is giving next: every member of the group interact with his/her partner to get the desired result in the robotic agent, but also the interactions of each team with the others must be coherent to get the desired behaviour of the robotic swarm.

Following, it is explained each component of the robotic swarm.

A. *Sensors*

Traditionally, the educational robotic systems for the school do not accept devices of other manufacturer. This fact makes difficult the access to spare parts and the design of new prototypes. On the contrary, TEAC^{2}H-RI is an open and modular architecture, i.e. the accessories that come with the kit can be improved or replaced. Even more, new accessories can be added, being therefore more adaptable to other technologies. These concepts are very important for the strengthening of the design, creativity and development skills in the user. The kit has five different sensors for each robotic agent. In these sensors the learner can experiment and modify the parameters of the equations that rule them. Also, additional pieces are provided to the learner, in order that he/she may develop new sensors. In Fig. 2 it can be seen the infrared sensor. In 2a the emitter of the infrared sensor does not have any resistance. In 2b a resistance of 10Ω is put on the emitter. In 2c the Velcro, used to hold the sensor to the robot, can be observed. In 2d the sensor is in the front side of the robot. Due to the fact that the resistance of the infrared sensor emitter can be changed, the maximum distance sensed can be modified. In this manner and based on the textbook, the concept of electrical resistance, current and voltage can be understood.

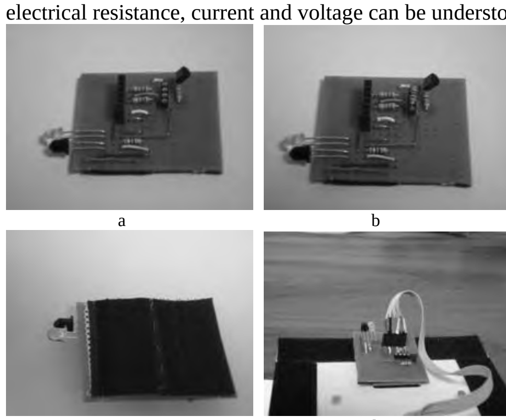


Fig. 2 Infrared Sensor a) without resistance b) with a 10Ω resistance c) with the Velcro at the bottom side d) in the front side of the robot

In the kit each robot has infrared sensors to avoid obstacles or to follow tracks, photoresistance to recognize day and night, touch sensor to avoid obstacles, sound sensor to detect claps, and also it can be added other sensors including, temperature sensor using NTC resistance and ultrasonic range sensor.

B. Mechanical Design

The chassis of both robots is very similar; its size is 23cm long, 14cm wide and 10cm high. The chassis has a first level where the motors card is and a second level where the Main Board is. In the outline of both levels, it is glued male Velcro. It allows changing the position of each sensor and explores the operation in function of the position. The design of the chassis, which could be assembled or disassembled, is ideal for experimentation and understanding of physical and mathematical concepts. For example, when it is modified the distance between the motors, the position of the sensors or the distribution of the mass in the platform , it could be understood some concepts, including centre of mass, torque and beam interference. An additional advantage of the mechanical design of TEAC^{2}H-RI is that its components are very cheap and easy to acquire in Colombian market. Fig. 3, shows the mechanical construction of the chassis of one TEAC^{2}H-RI robot.

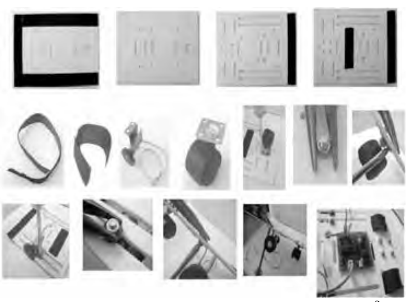


Fig. 3. First construction stage of the chassis of one TEAC^{2}H-RI robot.

C. Programming Language

It will be used the programming language CodeWarrior of the Metrowerks Company, which is supported in C. However, we are considering developing an easier programming language to the user, which not required previous knowledge in algorithms and give an introduction to this subject. [20] [21]

D. Textbook

The textbook contains the topics of physics, mathematics, electronics, algorithms, and artificial intelligence that will be addressed in the course. This is one of the most important issues, which has to be handled carefully, because require a language, concepts and adequate examples, to catch the attention of the learners. If it is not regarded with great caution, the reverse effect can be caused, i.e. rejection or phobia for the technological topics [6]. The figure 3 was taken from the textbook.

VI. Conclusion and Future Work

The production cost of each robot is around \$65US, which is very attractive in comparison with other options in the market. This fact facilitates the expansion of the educational robotics to a wider number of regions.

The implementation of educational robotics in developing countries must be done with a previous reading of the economical and social context. It is important to select the best tool that fit the reality. In our project, this reading was done and a new educational tool was developed: TEAC^{2}H-RI.

The open architecture of TEAC^{2}H-RI allows implementing new devices in the robot. This fact is very important because the learner would like to improve the performance of his/her robot through the design and development of new sensors or actuators. This characteristic, strengthen the creativity and design skills.

The proposed methodology is related to the steps to follow in a research project. In this way, implementing this methodology could be strengthened the research skills of the learner. It is extremely important in the developing countries, like Colombia.

In this first approach, the social group chose to work on is the Senior High School students from the Metropolitan Area of Valle de Aburra – Antioquia – Colombia. However, future works will cover not only the Metropolitan Area but also other Colombian and Latin American places.

In upcoming projects will be developed new programming languages with an easier interaction of the learners with the educational tool TEAC^{2}H-RI. Additionally, it will be brought an introduction in the programming world.

Currently, there is a human-robot interaction but it is very weak and it is necessary to consider the addition of new hardware and software devices which must adapt the kit to the user necessities [19].

The kit TEAC^{2}H-RI should not be taken as the panacea for the educational problems in Colombia, neither in the developing countries, because it is necessary the collaborative work between Government, School, University and Private Sector to evolve to the suitable educational model [5, 3], in which the teacher will give a greater commitment and effort but also will receive better fruits.

References

[1] Zuckerman O., "Historical Overview and Classification of Traditional and Digital Learning Objects". MIT Media Laboratory, 2006

[2] Piaget, J. Inhelder, B. 1966. La psychologie de L'enfant. Paris. P.U.F.

[3] Papert S. 1990. A Critique of Technocentrism in Thinking about the School of the Future.1990 M.I.T. Media Lab Epistemology and Learning Memo No. 2.

[4] Papert S. 1980. The Gears of my Childhood. Mindstorms: Children, Computers, and Powerful Ideas. Basic Books, New York.

[5] UN Millennium Project 2005. Innovation: Applying Knowledge in development. Task Force on Science, Technology, and Innovation.

[6] Chiou A. 2004. Teaching Technology Using Educational Robotics. En: Scholarly Inquiry in Flexible Science Teaching and Learning Symposium. (Sydney, Australia, Octubre 1, de 2004)
[7] Anderson M., Thaete L, Wiegand N. Player/Stage: A Unifying Paradigm to Improve Robotics Education Delivery. RSS Workshop on Research on Robotics in Education 2007 (Georgia, USA, June 30 , 2007)
[8] Barrera P., Martin F., Robles G. Cañas J., Matellan V. 2005. Modelos y herramientas de software libre en el campo de la robótica. En las memorias del congreso Tecnologías del Software Libre (La Coruña, España, Julio 7-8, 2005)
[9] Martin F. 2001. Robotic Explorations. A hands-On Introduction to Engineering. New Jersey Prentice Hall. 2001.
[10] Blank D. 2006 Robots Make Computer Science Personal. In Journal of Communications of the ACM. Vol. 49. Ed. ACM Press, New York, NY, 25-27.
[11] Fischertechnik. http://www.fischertechnik.de/en/index.aspx. Fecha de acceso: 09 de October de 2009
[12] UNESCO Institute for Statistics, 2006
[13] Dias M., Browning B., Mills-Tettey A., Amanquah N. (2007) Robotics Education in Emerging Technology Regions. En AAAI Spring Symposium: Robots and Robot Venues: Resources for AI Education. (Menlo Park, USA, 26-28 Marzo 2007)
[14] Urrea C. 2001. Designing Robotic Artifacts: Rural School-Community Collaboration for Learning. 2001 Media Lab MIT.
[15] Zuckerman O., Resnick M. 2003. System Blocks: A Physical Interface for System Dynamics Learning. MIT Media Laboratory. 2003.
[16] Ausubel, D.P. 1963. The Psychology of Meaningful Verbal Learning: An Introduction to School Learning, Grune & Stratton Inc., New York, 1963 Bureau of Educational Research, University of Illinois. Grune & Stratton, Inc.
[17] McKerrow P. Robotics research and education with LabVIEW. In Proceedings of the Apple University Consortium Conference (New South Wales, Australia, Abril 25-28, 2000)
[18] Acuña A. 2006. Projects for Educational Robotics: Engines for the Innovation. In the proceedings of the Fourth International Conference on Multimedia and Information and Communication Technologies in Education. Current Developments in Technology-Assisted Education (Sevilla, España, Noviembre 22-25, 2006)
[19] T. Fong, I. Nourbakhsh, and K. Dautenhahn, “A survey of socially interactive robots,” Robotics and Autonomous Systems vol. 42, no. 3–4, pp. 143–166, 2003
[20] Piperidis, S. Doitsidis, L. Anastasopoulos, C. Tsourveloudis, N.C. 2007. A Low Cost Modular Robot Vehicle Design for Research and Education In proceeding of the 15th
[21] Wedeward K., Bruder S., Sarkodie-Gyan T. 2001. Exploration in Robotics: Outreach to Secondary Education. In Proceedings of the AAAI 2001 Spring Symposium Series. (California, USA, Marzo 26–28, 2001)

Selecting Patterns for the Evaluation of Systems

Dmitry Zub, Erki Eessaar
Tallinn University of Technology, Estonia
dmitri.zub@mail.ee, eessaar@staff.ttu.ee

Abstract **- Earlier, we proposed a general pattern-based evaluation method of systems. The method is based on the use of Analytical Hierarchy Process and the use of patterns as criteria. This paper describes a process of selecting patterns that would be used during the evaluation. This process is an important step in the method. Firstly, we describe the process, based on an example. More precisely, we select a set of usability patterns that one could use to evaluate e-learning systems. After that, we propose a generalized process for selecting patterns.**

I. INTRODUCTION

There exist several types of requirements for software systems. For example, functional requirements specify the functionality of the system while quality requirements show how well a system should perform its intended functions.

Performance is a very significant quality requirement. It indicates how well a system should complete its tasks (speed, accuracy and the resources it takes). Another important quality requirement is *usability*. It shows how easy is to learn to use a system, how efficient it is in daily use, and some other aspects.

In some types of software systems, like educational software, the usability is especially important. If students, instead of studying, have to wonder how computer software works, then it disturbs the learning process. We conducted a study at the Tallinn University of Technology (TUT) [1]. During the study, we compared a TUT in-house developed e-learning system with some internationally known e-learning systems. The goal of the project was to evaluate the TUT system against other systems and to give some practical recommendations about improving its usability. One of the results of this project was *a general evaluation method* of systems that was developed to repeat such evaluations in the future.

In this method, we used *patterns* as criteria for the evaluation of systems. In usability evaluation of e-learning systems [1, 2], the selection of patterns was based on the experience of an expert. The *first goal* of this paper is to explicitly define a procedure of selection of patterns. The *second goal* is to illustrate the selection process based on usability patterns and propose a set of usability patterns that one can use for the usability evaluation of e-learning systems.

The rest of the paper is organized as follows. Firstly, we introduce the basic concepts that are related to our research and shortly describe our pattern-based evaluation method. Next, we explain the selection process of patterns based on a concrete example. After that, we give an overview of related work. In the next section, we discuss our selection process and present a generalized procedure. Next, we describe possible directions for future work. Finally, we draw conclusions.

II. BASIC CONCEPTS

A. Usability

Usability requirements are non-functional requirements. According to [3] a system with good usability should (a) be easy to learn, (b) make it possible to complete common tasks quickly, (c) be well memorable, and (d) make it possible to recover from errors quickly.

While these factors explain what usability is all about, it is rather difficult to use them in practice. Many methods exist that help us to evaluate usability. Folmer and Bosch [4] propose a possible classification of these methods. According to them there are three types of methods. The purpose of *usability inquiry* methods is to get information about likes, dislikes, and needs of users by talking to them and observing them working. The examples of such methods are focus groups [5], interviews [3], and questionnaires [3]. In *usability testing* potential users complete some typical tasks by using the system. For example, in "Thinking aloud" method a user verbalizes his/her thoughts or opinions while interacting with the system [5]. In *usability inspection* methods usability experts (and sometimes other software developers, users, and experts in other areas) examine usability related aspects of the system. One of the best-known inspection methods is "Heuristic evaluation" [6, 7]. In this method usability experts examine the system that is being tested and decide whether the system corresponds to some evaluation criteria (heuristics). Schmettow and Niebuhr [8] propose a pattern-based usability inspection method, according to which patterns should be used as heuristics in usability inspections. We also use patterns in our evaluation method. Therefore, in the next section, we shortly introduce the concept of patterns.

B. Patterns

Christopher Alexander, one of the well-known proponents of the pattern concept, describes patterns [9]: "Pattern is a three-part rule, which expresses a relation between a certain context, a problem, and a solution." Initially this concept was used in architecture, but later it became popular in other areas like software design [10], database design [11], and pedagogy [12].

Several formats exist for writing down a pattern. However, there are some parts that are used in all the formats. Every pattern has the *name* and usually it provides some kind of hint to the solution of the problem, the description of which is in the pattern. The body of each pattern consists of at least three parts. First part describes *context* (that means background or

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_14,

situation) in which a problem can occur. Secondly, the *problem* itself is described. And finally, the *solution* to the problem is described together with discussion of how to implement this solution. Now, we describe our method and how it was used for the evaluation of usability of e-learning systems.

C. A Pattern-based Evaluation Method of Systems.

We use Analytic Hierarchy Process (AHP) [13] for the evaluation of systems. The general idea of our proposed method is to compare a system *s*, which we want to evaluate in terms of an overall goal *g*, with similar systems by using a set of patterns that correspond to *g*.

We have to create a hierarchical model to solve a complex evaluation problem, according to AHP. On the highest level of this model is the *goal*, which is the overall objective that we want to achieve. For instance, the goal may be to find relative usability of different e-learning systems.

At the next level of the model are *objectives*. They are the main aspects that we should take into account while working towards achieving our goal. These objectives may have some *sub-objectives* that are located at the next level of the model. We propose to use patterns as criteria to evaluate how well systems satisfy objectives.

The selected patterns must correspond to the overall goal of the model. In case of evaluating an e-learning system, we might use web-usability and e-learning patterns as criteria.

And finally, at the lowest level of the model, there are *alternatives*. They are objects, between which the choice will be made. We propose to use systems as alternatives. One of the systems must be a system that we want to evaluate.

So in general, we are evaluating which of the alternatives corresponds better to our objectives. AHP method uses pairwise comparisons where elements of lower level are compared with each other in relation to some element of the model that is located at the next higher level. Therefore, objectives are compared with each other in terms of the goal. Possible sub-objectives are compared pairwise in terms of the objective. Finally, alternatives are compared with each other against objectives (or sub-objectives). For example, in [1], we used an in-house developed e-learning system and in addition three other e-learning systems as alternatives.

The steps of our method are the following.

1. Selection of the overall goal *g* of the evaluation. Based on that we have to select criteria (patterns) and alternatives (systems).
2. Selection of systems (alternatives) for the evaluation. One of the alternatives must be the system *s* that we want to evaluate. Other alternatives must be systems that are similar to *s*. The first criterion for selecting a system, with which we want to compare *s*, is its availability. The evaluators must have access to the selected systems so that they can perform tests based on these systems (see steps 5–6). Forman and Selly [14] note that psychological as well as mathematical evidences suggest “when deriving priorities from paired comparisons, n, the number of factors being considered must be less or equal to nine”. So if there are more such possible alternatives (systems), then it is necessary to further limit the set of systems. One should select the systems that have high popularity or quality (in terms of the goal of the AHP model). For the selection of alternatives one could use the results of surveys, opinions of experts, and personal experiences. If one selects systems that have low quality according to experts, then the good result of evaluation means only that *s* has a good quality compared to the other selected systems. Our approach is not well usable if one wants to evaluate a groundbreakingly new system that has few or no similar systems.
3. Selection of patterns (criteria) for the evaluation. We discuss the selection of criteria in the next section.
4. Analysis of the patterns, which are selected for the evaluation, by comparing them pairwise in relation to the goal *g* of the evaluation. The result of the step is an AHP model that shows us the relative importance of each pattern in terms of *g*.
5. Writing of tests that are needed to evaluate the systems in terms of the selected patterns.
6. Performing the tests based on the selected patterns.
7. Analysis of the results of the tests. Pairwise comparison of the systems in terms of each pattern based on the results of tests.
8. Calculation of final weights for each system based on the relative importance of patterns. The resulting model shows us how well each selected system (including *s*) corresponds to the goal of evaluation.
9. Identification and solving of the main problems that were discovered during the testing. It means partial redesign and reimplementation of system *s*.
10. Retesting of *s* to find out whether it has improved.
11. Analysis of the patterns used in the evaluation.

III. SELECTION OF PATTERNS

In our previous study [1], the selection of patterns was based on the experience of an evaluator. In this paper, we explicitly explain the method for selecting patterns. Firstly, we introduce the method based on an example.

Initially, we have to select collection of patterns that we will use for the analysis. We select Web Usability (WU) pattern language [15] because all the systems under discussion in the study [1] are web-based systems and we used patterns from this language in [1]. Therefore, we can compare the selection based on the experience of an evaluator and the selection that is proposed in this paper.

One should take into account the following aspects during the selection of patterns.

- The number of patterns that will be used for the evaluation. If we use AHP method, then probably *nine* is the maximum number of patterns one could use (see section II) as criteria. We can use more patterns if the criteria have sub-criteria but it complicates the evaluation. On the other hand, WU language contains 79 patterns [15].
- Functionality of the system. If a pattern is about a function that is not necessary or important for the system,

which we want to evaluate, then we do not have to consider the pattern.

- A pattern *p* could be unsuitable for the evaluation because it is not important in terms of the overall goal of the analysis or the evaluators do not have enough information about systems to evaluate them in terms of *p*.
- Some patterns could be related to each other. For instance, they could be similar to each other. In this case, we should decide what patterns to choose from the set of related patterns.

The first step of pattern selection is the classification of patterns. Firstly, we should examine what information the selected pattern language already contains. Hence, we now describe the format and grouping of patterns in WU language.

A. Groups of Patterns

The patterns of WU language are described in a classical format close to the one that is used in [9] by Alexander. Along with traditional pattern information (like name, context etc) we could use information about related patterns and star ranking of patterns.

Patterns of WU language are also grouped in categories according to a possible order of their use. The groups range from "Getting started with the site" to "Dealing with workflow and security".

Unfortunately, the existing grouping of patterns of WU language is not very helpful. The reason is that patterns that could be of interest to us are located in different groups. Therefore, in the next section we divide the patterns into other groups based on the problem, the solution of which they specify.

B. A Classification of Web Usability Patterns

After examining the patterns and problems that they describe, we decided to group them based on the types of problems that the patterns help us to solve. This way it would be easier for us to select patterns for the evaluation. The groups (and the problem areas) are the following. (See table one for number of patterns in each group)

- Getting started – development process of systems (from setting of business objectives to testing).
- Effectiveness of use – problems with navigation, search, speed of using the system.
- Functional patterns – patterns that give recommendations about the functionality of systems and organization of functions within a system.
- Help patterns – the application of these patterns makes it easier for users to learn to use the system.
- Content patterns describe what information should be published at the site and how the text should be written.
- Design patterns in our case describe how to implement functions of the system, and how to organize the building-blocks of systems.
- Technical patterns concern such things like the use of HTML, security, and others.

These groups were quite roughly defined. Therefore, in our opinion some patterns could belong to more than one group.

The next step in the process of selecting patterns is to exclude patterns that one cannot use for the evaluation due to their unsuitability for evaluating already existing systems.

C. Excluding Patterns from the Evaluation

As we mentioned in the previous section, WU pattern language has some patterns that deal with the initial planning of systems (Group "Getting started"). Most of the patterns in this group are not suitable for our evaluation. In other groups, there are also some patterns that deal with rather technical issues that are hard (or impossible) to check in the evaluation process. For example, pattern "Cache transactions".

The only pattern, which could be useful for us from the set "Getting started", is "Two years old browser". The reason is that it describes not only how to test but also what to test. (Check whether the system works in computers with old hardware and with different web-browsers.) Hence, we propose to use this pattern in the evaluation.

D. Finding the Best Candidates from the Groups

The goal of this step is to determine for each remaining group one or more patterns that are the most important in the context of the evaluation.

Patterns About the Effectiveness of use of the System

In this group there are several patterns about the speed of using the system like "Download times" and "Content before graphics". We do not include these patterns in the evaluation. We assume that nowadays there are not many problems with internet speed and there is no need to download large files in e-learning systems. As you can see, selection of patterns requires assumptions about systems that we evaluate.

Search for information is very important task for all systems, not only educational ones. Hence, the majority of the patterns that describe how to organize navigation are important. Therefore, we discuss them next in more detail.

Navigational Patterns

Firstly, we shortly describe patterns (from [15]) that in our opinion cover navigational aspects.

"Site map" – Provide a single page that gives visual and textual representation or overview of the entire site.

"Search box" – Allow visitors to search from a site.

"Sense of location" – In case of each page, provide visitors clear clues about where they are on the page.

"Canonical location" – Make sure that the page name is displayed prominently

"Breadcrumbs" – Do not rely solely on breadcrumbs for navigation unless you are short of space.

"Site logo at top left" – Clicking on the logo is a way to go to the homepage.

"Navigation bar" – Provide a bar on the home page and possibly on other pages that allow users to go to the main sections of the site.

"Go back to safe place" – Allow users to mark the place they want to return if they are lost.

"Back button" – Do not disable back button of browsers. Design transactions the way that they can be undone.

If we look at the patterns described above, then we see that there are some patterns that could be grouped together. For example, "Search box" can be an addition to "Site-map". "Sense of location" is more general variant of "Canonical location". "Canonical location" provides a recommendation to use cross-links across site hierarchy. One can use the solutions from "Breadcrumbs", "Navigation bar", and "Site logo" patterns to create these links. "Back button" (or "Site logo") can be used as a variant of "Go back to safe place". Therefore, we think that from these patterns the most important are "Site-map", "Sense of location", and "Go back to safe place".

Content (and Text) Patterns

This group of patterns explains what information should be on a web-site and also how it should be represented. In our opinion, most of content patterns (like "Links to many sites" and "Human touch") are not very important for e-learning systems. From this group we add to the evaluation "Natural metaphor". If this pattern is used, then probably it is easier to learn to use the system and to understand it. In addition, "Home page" pattern could be used in a sense that it would be nice if links to all main features of the system would be available at home page.

There is a similar situation with the text patterns. Most of them provide recommendations about how text should be represented ("Design page for scanning", "Short text", and "Separate print pages"). That is surely important for learning materials, but not for the system itself. From that category of patterns, we include "Acceptable wording" for the evaluation. If the system uses well-known terms, then it is easier for users to understand the new system and subsequently learn to use it.

Technical Patterns (HTML, Security)

We found no suitable patterns in this group as all of them rather technical.

Functional Patterns

The most important functional pattern in the context of usability evaluation is probably "No unpleasant surprises" – system behavior should be predictable and reliable. Also "Exploit closure" can be suitable – this pattern means that before completing some main tasks it should be guaranteed that a user has completed all the necessary preliminary tasks.

Help Patterns

Help patterns are quite important from the point of view of usability and they improve the ease of learning and the effectiveness of use of a system. We have selected three patterns from that area. If a system gives a user a "feedback", then he/she understands better what is going on while working with the system. "Sense of progress" helps user to understand better the task he/she is completing. "Context sensitive help" enables a user to get necessary information about the task without browsing help documentation.

Design Patterns

There are a lot of design patterns according to our classification. Large part of them could be useful in our opinion because they describe how to improve different aspects of usability. Let us describe those patterns in more detail.

"Follow standards" and "Keep it simple" are rather abstract. However, if these two patterns are followed, then it would be easier to learn and to use a site.

"What you see is what you can use" – If a user sees on the site only things that he/she can use in the learning process and he/she is not distracted by something else, then it is also good for learning.

"Mandatory fields" – Avoid the use of mandatory fields, and while using them provide information about how to fill them in forms.

"Oblique Landmarks" – Indicate somehow whether some features of a system are usable indirectly or when a user is in a certain state.

"Button gravity" – Users usually expect "Submit" buttons at the bottom of the page. Therefore, you should locate them there (or make them more visible).

"Pipeline interaction" – Give users chance to revisit the earlier steps of the workflow and correct provided data if needed.

Selecting the Final Subset of Patterns

Based on the previous discussion, we have the following patterns left as candidates.

1. Two year old browser
2. Site-map
3. Sense of location
4. Go back to safe place
5. Natural metaphors
6. Home page
7. Acceptable wording
8. No unpleasant surprises
9. Exploit Closure
10. Feedback
11. Sense of progress
12. Context sensitive help
13. Keep it simple
14. Follow standards
15. What you see is what you can use
16. Mandatory fields
17. Oblique landmarks
18. Button gravity
19. Pipeline interaction

We found too many patterns (19) that can be used in the evaluation. How can we reduce the amount of patterns to *nine* [14]? Firstly, the list contains some patterns that specify a possible solution (or part of solution) to some other more general (abstract) pattern. For example, if we use pattern "Button gravity", then it means that we also use pattern "Follow standards". If users expect that "Submit" buttons are located at the bottom of the page, then it means they are used to that because of previous experience and it became a standard of

TABLE I

PATTERN GROUPS AND NUMBER OF PATTERNS

Groups of patterns	GS	Eff	Funct	Help	Cont	Des	Tech	In all
Number of patterns	11	21	8	5	10	20	13	88
Of them selected	1	3	2	3	3	7	0	19

some kind. So, we should remove the more specific pattern from the list of selected patterns and only use the more general pattern. According to this principle, we propose to eliminate pattern 11 (as it is a sub-pattern of 3), patterns 5 and 7 (and keep pattern 13), keep pattern 17 and eliminate pattern 15. In addition, we keep pattern 14 and eliminate pattern 18. Pattern 19 is a sub-pattern of 11, which in its turn sub-pattern of 3. This way, we have eliminated six patterns (5, 7, 11, 15, 18, 19) and decided to keep four patterns (3, 13, 14, 17).

The second way to decrease the number of patterns is to look through all of them once again and to decide if some of the patterns are really important for the systems, which we want to evaluate. For example, while we think that "Oblique Landmarks" pattern increases the usability of systems, it is possible that in e-learning system there are no functions that are available only to some special types of users. "Home page" pattern states that we should draw the attention of potential users to a web-site with a home page. That is not important for learning. The requirement that all the important functions should be visible at home page can in some way satisfied with "Site-map" pattern. "Exploit closure" pattern concentrates more on how system should be designed so that all the tasks would be completed in the right order. "Sense of location" should provide users necessary information so that it would be clear in what order tasks should be completed. Finally, "Context sensitive help" is quite important because if a user needs to understand how a function works, then he/she can get the necessary information without searching that in help files. A recommendation to give information about filling "Mandatory fields" in forms is similar to the "Context sensitive help", so we do not include "Mandatory fields" in the final subset of patterns.

At the end, we have selected patterns 1–4, 8, 10, 12–14 for the evaluation. In the next section, we compare this selection with the one that was made in our earlier research [1].

E. Comparison of the two Selections.

In our previous research [1] two types of patterns were used for the evaluation: web-usability patterns and e-learning patterns. Therefore, five patterns of each type were selected. For the web-usability, the following patterns were selected. "Two year old browser", "Site-map", "Sense of location", "Feedback", and "Follow standards". All these patterns were selected in the current research as well. It means that the choice of patterns in [1] was good and repeatable. But it was based solely on the experience of one expert and the process of making this choice was not explicitly explained.

IV. RELATED WORKS

As it was already mentioned in Section II, a pattern describes "a solution to a problem" so that the solution could be used over and over again in different contexts. There exist collections of patterns (pattern languages) that describe problems and solutions of a specific domain. Examples from the field of usability are Wu Language [15] and collections of van Welie [16] and Tidwell [17].

There exist some examples of the use of patterns for the evaluation of systems. Kramberg [18] examines how much IBM WebSphere software supports the use of control-flow patterns from [19]. Kramberg [18] does not select a subset of patterns for the evaluation. There exists a proposal [8] to use patterns in the usability inspection. Unfortunately the method of selecting patterns is not clearly defined. Moynihan et al. [20] describe a prototype of an expert system for the selection of software design patterns. But the goal of this system is "leading a designer to a suitable design pattern to the problem at hand" [20] and so it cannot be used to evaluate a system.

During our selection process, we tried to classify the patterns. The idea of classifying patterns is not new. One of the reasons for classifying patterns is the big amount of pattern collections in the literature. Henninger and Correa [21] state for instance that there exist 14 collections of user interface patterns containing 425 patterns in total. Hence, some papers discuss how patterns and pattern languages can be organized. Van Welie and van der Veer [22] discuss how pattern languages in interaction design can be structured. There are attempts, like [23], to create repositories of patterns. One of the problems of storing many patterns in one repository is that different authors may use somewhat different formats for describing patterns. In order to solve this problem in Human Computer Interaction, a unified pattern format was developed named Pattern Language Markup Language (PLML) [24].

V. A GENERALIZED PROCEDURE

We have developed a method for evaluating systems by using patterns as the criteria. In [1, 2], we demonstrated how to use the method to evaluate usability of e-learning systems based on usability patterns. In [1], we selected patterns for the evaluation based on the personal experience and knowledge of an evaluator without trying to explicitly explain the process of selection. In this study, we repeated the selection process and divided it into steps so that it would be possible to reason about it. The process itself consists of *four* steps.

Step 1: Selection of one or more pattern languages (sets of patterns) or individual patterns. The selected patterns form the initial set *S1* of patterns, from which we select the final set of patterns that will be used for the evaluation.

Step 2: One has to look through all the patterns in *S1* and divide them into groups. All the patterns in one group must help us to solve similar problems.

Step 3: Some patterns might be unsuitable for the evaluation process (at least for the evaluation of existing systems). One has to eliminate these patterns from *S1*.

Step 4: From each group (step 2) one has to select a subset of patterns. The selected patterns are the main criteria and form set *S2*. The cardinality of *S2* must not be bigger than *nine*. Selection must be based on the suitability of the patterns to the systems in the analysis. For example, while "Internationalization" is important for the sites of organizations that operate in several countries, it is not so important for an e-learning system where learning is going on in one language.

In our opinion a pattern *p*, which defines a general rule, is very suitable for the evaluation if there are one or more "sub-patterns" that can describe specific (variants) of solution for a problem defined in *p*. We can use this principle also for selecting a subset of patterns from each group (step 4). If a selected pattern *p* has a set of specific sub-patterns, then it is also possible to use these patterns as sub-criteria of the main criterion (*p*) in the AHP analysis. The amount of *sub-criteria* of a criterion should also be no more than *nine*, because we have to compare the sub-criteria pairwise.

The proposed process still needs decisions of experts who perform evaluation. However, the procedure should facilitate the work of evaluators. If decisions, which are made during the selection, are explicitly documented, then it also allows interested parties to better understand the evaluation results.

To get the final set of patterns for the evaluation, we can repeat step 4 once more to compare patterns from different groups. The pattern-based evaluation method is not well usable if one wants to evaluate an aspect *a* of systems, and *a* has few corresponding patterns that specify the best practices.

To facilitate the process of pattern selection, the following additional information about each pattern should be recorded in the repository of patterns: (a) a classification of patterns based on the problems the patterns deal with; (b) is this pattern suitable for the evaluation process?; (c) the history of using the pattern for the evaluation of systems; (d) does this pattern provide a general advice or some practical solution?; (e) what is the rating of the pattern?; (f) how is this pattern connected to other patterns in the collection?

VI. Future Work

The future work in this area includes the creation of a system, which supports the evaluation process. Among other things it should support the management of patterns and all the necessary information for the selection (see Section V) of patterns. Another direction of research is to study the following aspects of pattern-based testing process.

- What kind of tests to use in the evaluation?
- How many patterns to evaluate in one test? Should a test cover one pattern or a set of them? This information could be later stored in a repository so that the future use of this test would be easier.
- Who will be the testers of the system and how many people should participate?
- How to determine the relative importance of the problems discovered during the testing, and how to classify them?

VII. Conclusions

We have proposed an evaluation method of systems that is based on Analytical Hierarchy Process (AHP) and patterns. Patterns are criteria, based on which is possible to evaluate how well systems correspond to different objectives. Psychological and mathematical evidences suggest that the amount of criteria (that correspond to the objectives), that one can use in AHP, should not be bigger than nine. For instance, WU language of usability patterns contains 79 patterns. Therefore, it must be possible to select a set of patterns for the evaluation from a bigger set of patterns.

In this paper, we proposed a systematic and structured process how to select a set of patterns that one could use to perform evaluations of systems. We illustrated the process by selecting a set of usability patterns that one can use to evaluate e-learning systems.

References

[1] D. Zub and E. Eessaar, "Pattern-Based Usability Evaluation of E-Learning Systems," International Conference on Engineering Education, Instructional Technology, Assessment, and E-learning (EIAE 07), pp. 117–122, Springer, 2008.
[2] D. Zub and E. Eessaar, "Pattern-Based Usability Analysis: Comparison with a Case-Study," International Conference on Engineering Education, Instructional Technology, Assessment, and E-learning (EIAE 08), Springer, 2009, in press.
[3] J. Nielsen, "The Usability Engineering Life Cycle," *Computer*, vol. 25, pp. 12–22, 1992.
[4] E. Folmer and J. Bosch, "Architecting for Usability: A Survey," *The Journal of Systems and Software*, vol. 70, pp. 61–78, 2004.
[5] J. Kontio, L. Lehtola, and J. Bragge, "Using the Focus Group Method in Software Engineering: Obtaining Practitioner and User Experiences," International Symposium on Empirical Software Engineering, pp. 271–280, 2004.
[6] J. Nielsen, *Usability Engineering*. Academic Press, 1993.
[7] J. Nielsen, "Heuristic Evaluation," In: Nielsen J., Mack R. L. (eds.) Usability Inspection Methods. New York: John Wiley & Sons, 1994, pp. 25–62.
[8] M. Schmettow and S. Niebuhr, "A Pattern-based Usability Inspection Method: First Empirical Performance Measures and Future Issues," HCI 2007, vol. 2 of People and Computers, pp. 99–102, 2007.
[9] C. Alexander, *The Timeless Way of Building*. New York: Oxford University Press, 1979.
[10] E. Gamma, R. Helm, R. Johnson, and J.O. Vlissides, *Design Patterns – Elements of Reusable Object-Oriented Software*, Addison-Wesley, 1994.
[11] D.C. Hay, *Data Model Patterns: Conventions of Thought*. New York: Dorset House Publishing Company, 1996.
[12] J. Bergin "Fourteen Pedagogical Patterns," Fifth European Conference on Pattern Languages of Programs, 2000.
[13] T.L. Saaty, "How to Make a Decision: The Analytic Hierarchy Process," *European Journal of Operational Research*, vol. 24, pp. 19–43, 1990.
[14] E. H. Forman and M.A. Selly, *Decision by Objectives*. World Scientific Publishing Company, 2002.
[15] Web Usability Patterns. Retrieved September 25, 2009, from http://www.trireme.com/WU/home.htm
[16] M. van Welie, Interaction Design Patterns Library. Retrieved September 25, 2009, from http://www.welie.com/patterns/
[17] J. Tidwell, Designing Interfaces. Retrieved September 25, 2009, from http://designinginterfaces.com/
[18] V. Kramberg, "Pattern-Based Evaluation of IBM WebSphere," BPEL Master's Thesis, Universität Stuttgart, 2006.
[19] V.M.P. van der Aalst, A.H.M ter Hofstede, B. Kiepuszewski and A.P. Barros, "Workflow patterns," *Distributed and Parallel Databases*, vol. 14, pp 5–51, July 2003.
[20] G.P. Moynihan, A. Suki, D.J. Fonseca, "An Expert System for the Selection of Software Design Patterns," *Expert Systems*, vol. 23, pp. 39–52, 2006.
[21] S. Henninger and V. Correa, "Software Pattern Communities: Current Practices and Challenges," 14th Conference on Pattern Languages of Programs, 2007.
[22] M. van Welie and G. C. van der Veer, "Pattern Languages in Interaction Design: Structure and Organization," Interact'03, pp. 527–534, 2003
[23] A. Gaffar, D. Sinnig, H. Javahery, and A. Seffah, "MOUDIL: A Comprehensive Framework for Disseminating and Sharing HCI Patterns," CHI 2003 workshop on HCI Patterns: Concepts and Tools, 2003.
[24] S. Fincher, "Perspectives on HCI Patterns: concepts and tools (introducing PLML)," *Interfaces*, vol. 56, pp. 26–28, 2003.

Effectiveness of Collaborative Learning in Teaching Information Systems

Ioana Tuugalei Chan Mow
National University of Samoa

***Abstract*: This paper is based on the second phase of an investigation which explored student attitudes and experiences in collaborative and individualised learning in computer based instruction. It attempted to answer the questions a) whether Samoan students prefer learning in a cooperative learning environment as opposed to individualised learning and secondly b) whether collaborative learning is more effective than individualised learning? Research results indicated that students enjoy and show positive attitudes towards collaborative learning and that they perform well in a collaborative environment provided the essential conditions for effective collaborative learning are satisfied or present.**

I INTRODUCTION

This paper describes the second phase of an investigation which explored student attitudes and experiences in collaborative and individualised learning in computer based instruction. It attempted to answer the questions

- ***Do Samoan students prefer learning in a cooperative learning environment as opposed to individualised learning?***
- ***Is collaborative learning more effective than individualised learning*?**

The Computing Department at the National University of Samoa offers courses in Computer Science and Information systems. One of the challenges in teaching Information Systems is the provision of students with skills in developing real life systems in the classroom. In the workplace, these systems are developed using the team approach.Therefore, the use of collaborative learning in the learning environment for Information systems coordinates well with the team approach to Information systems design in real world applications.

Another motive for undertaking this study was the interest in the cultural effects of collaborative learning in the field of computer education in Samoa. Samoan culture is communal and based on communal practices[1]. Family members cooperate and share in all manner of responsibilities ranging from domestic tasks to maintaining family lands and providing for the welfare of the family.

A major drawback in Educational computing in Samoa has been the high cost of computer technology, manifesting itself in the shortage of computers and the high student to computer ratio. Enrolment numbers for computer courses have always been restricted by the availability of computers. If collaborative learning could be proven effective in teaching computing subjects, this would have a direct beneficial bearing on resource implications.

The current research was conducted in two phases. The research was structured such that the first phase initiated a general comparative investigation of the attitudes and performance of students in collaborative learning as compared to individualised learning. This initial phase spanned three weeks. The second phase provided a semester long and more in-depth study of the essential conditions necessary for optimal collaborative learning.

II METHODOLOGY

The research design used in the study was quasi-experimental, where the independent variable (learning environment) was manipulated and subsequent changes in the dependent variables (student attitudes to learning environment and student performance) were investigated. The quasi-experimental design was chosen for both phases in the study, as its strict control of variables enabled a distinct focus on causal relationships between learning environment and student attitudes to learning.

The population for the second phase was a 'captive' one consisting of students in a Computer Information systems class. The students in this phase had already completed an Introduction to Information Systems class in Information systems analysis and design. Phase 2 involved a much smaller group of students (9 students) and group selection was on ability so that each group of three was of mixed ability (high, low and average). The composition of the group is an important factor in CL and research in CL recommend heterogeneous groups as some difference in view points is required to initiate interactions between group members [2]. Ability was based on previous grades in CSB 283, the prerequisite class to this Information Systems class.

III PROCEDURES

Phase II of the research was conducted on students in an Information systems and database design class at the local university. The class was divided into heterogeneous groups of threes according to ability (high, average, and low ability). Selection and composition of the groups were decided by the lecturer, who was also the principal researcher.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_15, © Springer Science+Business Media B.V. 2010

To provide the context rich environment of the real world, the groups were assigned two real life projects. The first project involved the development of a computerised system to automate the checkin/checkout transactions of the university library. The second project involved the development of a computerised invoice entry system for a local water company.
The major activities in the two projects involved

- the groups carrying out investigative interviews on the clients to elicit such things as user requirements, problems in existing systems leading to a need for an improved system and also the feasibility of the new system.
- Systems analysis using a CASE software tool called ASCENT
- Systems design using ASCENT and Word for Windows for prototyping design of menus and user interfaces
- Implementation of the system using Microsoft Access.
- Testing of the database system and making the necessary revisions and improvements.

Since the two systems were planned and designed using Systems Development Life cycle approach, which involved several phases, assessment of the projects was continuous. For example, the deliverables for the early planning and investigative phases were assessed and feedback given on areas needing improvement before proceeding to the systems analysis and design phases and eventually to the implementation.
Throughout the two projects, which spanned the duration of a semester, students were expected to carry out planning, analysis and problem solving by means of extensive group discussions.

IV OPERATIONALISATION OF VARIABLES

In this study, the independent variable was the learning environment which had two values in its domain, collaborative learning and individualised learning. The dependent variables, were student performance and student attitudes to the learning environment. In the course of the study, the independent variable (learning environment) was manipulated and subsequent changes in the dependent variables (student attitudes to learning environment and student performance) were gauged using questionnaires, surveys and pre and post-tests.

A *Student performance*

In Phase II, the effectiveness of the learning environment is student performance measured or assessed as a group using criteria such as group presentations, grades for deliverables such as the Initial Data Model, Initial Process Model, Feasibility study, database tables, forms, menus and reports. The marks for the collaborative learning sessions were compared to marks from the previous Information systems class and the difference between the two sets of grades, analysed statistically, for significance.

Evaluation of group projects involved one group presentation for the Library system and written documentation such as data flow diagrams, entity relationship diagrams, data dictionary, user manuals and a disk copy of the final database system. Assessment was done on a group basis with one mark given to the whole group. Assessment from the two projects accounted for 80% of the overall assessment. The remaining 20% came from a 1-hour test which tested their knowledge and comprehension of database theory.

B *Student attitudes*

Student preferences for a particular learning environment and their feelings towards such an environment were measured by self-administered questionnaires, structured interviews and observations by the principal researcher. The questionnaire focused on issues of, the enjoyment of learning, self-esteem, motivation, learner control, and social interactions. For the purpose of the current study, the enjoyment of learning was interpreted and defined as students' valuation and appreciation of the learning experience.

V DATA GATHERING TECHNIQUES

Student attitudes to the learning environment were measured by self- administered questionnaires and also by observations by the researcher. In addition to the self-administered questionnaires, groups in Phase II were also interviewed on their views and experiences with collaborative learning.

VI RESULTS

The aim of the current research was to establish from the results of the study, the following propositions. Firstly, the study attempted to confirm that students in Samoa enjoyed collaborative learning and preferred it to individualised learning. Secondly, the study attempted to establish that students in Samoa performed better in a collaborative environment.
Hence, for purposes of the discussion, the main areas of focus were student performance and student attitudes to learning.

A *Student attitudes to Learning*

Student attitudes to collaborative learning were gauged from responses to the questionnaire administered at the end of the session and also from interviews with the researcher. Results from the various questionnaires and interviews, showed that the majority of the students viewed collaborative learning as an effective instructional strategy with positive effects on problem solving, self esteem, motivation and enjoyment. However, despite these outcomes, the results did not show that students preferred collaborative learning over individualised learning as the analysis of Phase 1 results did not reveal any significant differences in attitudes between the two treatments. Discussion with the students in the interviews revealed the following:

Question 1 Did you enjoy working together as a team?
Of the three groups in this second phase, the members of two of the groups, enjoyed working together as a team. Two of the members of Team 2 in the early stages of the exercise, initially resented working in a collaborative mode, due to the inability to resolve conflicts between group members. However, their attitudes changed as the exercise progressed.

Question 2 What problems, if any, did you have working together as a group?
Members in Group 2 indicated that they had had problems in scheduling times in which everyone would be available for group discussions. There were also feelings of resentment where some members felt that members in the group were not contributing their share of the work. Most of the members in the other two groups, felt that they had progressed well and had not had problems working as a team.

Question 3 Did everyone contribute equally or did some dominate the group?
The general feeling in Group1 and Group3 was that everyone contributed equally to the group. In Group 1, Student 1A and Student 1B were Accounting experts while Student 1C provided the expertise in computing applications, so there was some form of division of labour, as the members complemented each other's skills and shared their expert knowledge.

Question 4 Did working together help the student's understanding of the project?
All of the students agreed that collaborative learning improved their understanding of the project. There were aspects of the project one would not have thought of, had they not been told of, by other members of the group. Discussion of some of the issues in the project, led to awareness of further issues in the project that they, as a group, had to resolve and address. However, Student 2D and Student 2E in Group2, felt that working as a group actually led to confusion of issues in their group. When asked to elaborate, they explained that both members were convinced that their own views of the way in which some aspect of the project should be approached were the correct one and refused to accept any other perspective.

Question 5 Do you feel that collaborative training helps in pooling your knowledge and experience?
When asked about the value of collaborative learning in pooling their knowledge and experience, all of the members in Group 1 and Group3 felt that they would not have been able to develop and implement the systems without the help from fellow members and the lecturer.
Student 2D and 2E initially had doubts about the worth of group learning, but as the research progressed, they changed their opinions in favour of collaborative learning.

Question 6 Did you get helpful feedback in your group?
All of the students participating in the collaborative mode, indicated that they received helpful feedback from their group as the group members discussed the various issues in the project.

Question 7 Did working in a group help stimulate thinking?
As revealed in the interviews, all of the participants, agreed that group learning stimulated their thinking, in that the group discussions gave them new perspectives on the subject area and brought up other valuable issues related to the topic under discussion.

Question 8 Did working in a group give you new perspectives on the subject area?
All of the students agreed on the value of group learning on broadening one's perspectives.

VII DISCUSSION

The aim of the current research was to establish from the results of the study, the following propositions. Firstly, the study attempted to confirm that students in Samoa enjoyed collaborative learning and preferred it to individualised learning. Secondly, the study attempted to establish that students in Samoa performed better in a collaborative environment. Hence, for purposes of the discussion, the main areas of focus were student attitudes to learning and student performance. Student attitudes to Collaborative learning were both positive and negative.

i) *Positive attitudes to Collaborative learning*
The groups in which everyone participated and contributed to team progress enjoyed the experience and found that collaborative learning led to increased social interactions between the members of the group. As evident in all of the student interviews, students felt that solving problems and carrying out tasks in groups was a lot more effective than doing it individually. From student responses in student interviews, it was clear that practically all of the students in the collaborative group felt that it was an effective instructional strategy for the following reasons :

- *Collaborative learning taught students teamwork*

Students felt that collaborative learning taught them to work together as a team and share both the work-load and ideas. This was consistent with the findings of researchers such as [3] and [4] that demonstrated the positive effects of collaborative learning in fostering team-work.

- *Students enjoyed Collaborative learning*

Practically all of the participants, claimed, they had enjoyed working in the collaborative mode. This was especially true in groups where members had successfully worked together as a group and enjoyment was probably due to feelings of success. It could also be argued that enjoyment led to an increase in intrinsic motivation. These findings were consistent with the literature and it could be argued that these feelings of enjoyment were due to increased motivation and increased social interactions. Collaborative learning, as [5] pointed out was a social event where learning was fostered through peer interactions and social discourse.

- *Collaborative learning led to increased self confidence*

The positive effects of collaborative learning on self confidence and esteem have been established by findings of studies by [7] and [9]. This was confirmed by responses of the majority of the participants in the collaborative treatment, who maintained, that collaborative learning had boosted their self-confidence. Again this was evident mostly in groups where the collaborative mode had been successful leading to feelings of increased self-confidence. Furthermore, the mere fact that most of the decisions were the product of group discussions and consensus, meant that students felt a lot more confident in any decision making and tasks done by the group as it had the approval and confirmation of other people.

- *Collaborative learning promoted sharing of ideas*

Practically all of the students in the collaborative group, confirmed, that they felt that the benefits of collaborative learning lay in the ability to share ideas through group processing.

- *Encouraged discussion of issues from various view points enabling an in-depth probing of problem, task or concept.*

As pointed out by [5], knowledge is a social construct and the learning process is facilitated by social interactions amongst students as they collaborate with each other to actively construct their learning. In the collaborative learning groups, the sharing of ideas and varying viewpoints through discussions led to a deeper understanding and analysis of a problem, task or concept as it exposed the student to concepts they had not thought about before. These supported the findings of studies by [3], [4], [6], [7], [8] and many others, which showed that collaborative learning led to increased cognitive analysis and enhanced problem solving in learning.

- *More able students provided 'scaffolding' for the weaker ability ones*

The concept, referred to by [10] as the "zone of proximal development", that the more able students provide mentoring for the less able, was quite noticeable in the successful groups in Phase II. In these groups, the more able students helped the weaker students in understanding concepts. A direct result of this mentoring was the obvious improvement in self-confidence in the less able students and, more importantly the significant improvement in achievement of the weaker ability students in Phase II.

- *Increased motivation*

The students also indicated that collaborative learning promoted increased feelings of motivation, and this was particularly true in cases where collaborative learning had been implemented successfully. These findings are in tune with the findings of [3], and [8].

ii) *Negative attitudes to Collaborative learning*

It should be acknowledged, that student attitudes to collaborative learning in the current study, were not all positive. The research results also included issues in collaborative learning, which the participants found troublesome and aspects of the learning environment, which they found disagreeable. However, in viewing all of the student feedback, these negative aspects were quite minimal (less than 10 %) when compared to the overwhelming volume of positive feedback.

- *Lack of cooperation*

Perhaps the most common complaint raised by the participants was the lack of cooperation from some team members who were not doing their "*fair share*" of the work. The lack of cooperation also resulted in the lack of group processing, which, as recommended by key researchers ([3]; [11]) in the field, was one of the essential factors for successful collaborative learning.

- *Slow pace of work*

Another complaint voiced by the students in the collaborative mode was the slow pace of progress. It could be explained, that apart from some students wasting time on "off task activities", the longer time spent on tasks, was mainly due to more time spent on discussions and increased interactions amongst group members. However, to this end, the author would like to claim that this slow pace was actually beneficial, as the group members needed this time to carry out in-depth discussions, to critically reflect on the group's progress and plan of action and to resolve any conflicts and controversy which may have arisen. These activities, if carried out effectively, would have resulted in improved problem solving and promoted critical thinking within the group.

- *Resolving controversy*

Another issue, brought to the fore by the author's observations, was that of resolving conflict and controversy amongst group members.

According to [12],

> "Conflict exists whenever incompatible activities occur..."

Classification of conflicts include "controversy, which occurs when a person's ideas or opinions are incompatible with others" ([12]), and there is a need to arrive at some consensus or agreement. [12] have argued that controversy, when managed constructively, could lead to enhanced problem solving, enhance the quality of reasoning and decision making and strengthen liking, respect and trust amongst members.

In Phase II, the feelings of hostility between Student 2A and Student 2B, were due to their inability to resolve conflicts within the group. These students had opposing views on most issues. However, neither was prepared to come to some agreement or work out a settlement. Controversy, when resolved constructively would prove beneficial to collaborative learning. However, controversy, when unresolved would lead to resentment and a breakdown of collaborative learning. Inspection of achievement of these students, showed them up as having the least improvement. In fact, for student 2A, the effect of collaborative learning on achievement was negative. The inability to resolve conflicts led to a lack of group processing, and hence resulted in ineffective collaborative learning.

iii) Collaborative learning versus individualised learning
On the overall, the study confirmed the social benefits of collaborative learning such as enhanced self-esteem, enhanced problem solving skills, and the enjoyment of learning. These findings were consistent with those chronicled in the research literature on collaborative learning.

iv)The effect of communal culture on attitudes to collaborative learning
It could be argued that students enjoyed and showed positive feelings towards collaborative learning because they were already familiar with such practices in everyday life. In the 'faa Samoa' or the Samoan way of life, most of the activities or practices are done collaboratively or collectively. The group rather than the individual is important and groups can be the nuclear family unit, the extended family, the church group or the village to which one belongs. Therefore, it could be argued that students learning in a collaborative setting enjoy it as they are already enmeshed in these collaborative networks in their way of life and is an extension of their daily communal practices.

B *Student performance*
Students in Phase II showed a marked improvement, overall, in their academic performance, and this was consistent with the findings of the majority of studies in collaborative learning ([13]; [3]; [4]; [14]. It could be argued, that the success of collaborative learning in Phase II, was due to the presence of all the key elements, necessary for effective collaborative learning. The presence of individual accountability and feelings of personal responsibility, amongst group members, ensured the successful implementation of collaborative learning, resulting in the positive effects on student achievement. Within the groups, there were extensive discussions as members worked together as a team to collectively devise strategies for designing successful systems. Group members also took time aside to critically reflect and discuss the progress of the group. In fact, collaborative learning had fostered improved social relations and friendships amongst these group members.
It was evident then, from the discussions set out above, that for successful collaborative learning, certain conditions were essential and needed to be systematically incorporated into the educational process. These conditions include a commitment amongst members to work as a team, group processing, positive interdependence and individual accountability.

VIII CONCLUSION

The results of Phase II of the present study confirmed that students in Samoa perform well in a collaborative environment, provided that the essential factors for effective collaborative learning, such as positive interdependence, individual accountability, team and social skills and group processing, were incorporated into the educational environment.
To conclude, it is clear from the results of the research that if collaborative learning is to be successfully incorporated into the education process, further investigation is necessary. With the right conditions, and using the right procedures, collaborative learning could prove to be an effective inexpensive teaching strategy that could contribute positively to the education process in Samoa.

Bibliography

[1] Ochs, E. (1988). Culture and Language Development: Language Acqusition and language socialization in a Samoan village. Cambridge Press, p 71 – 78.

[2] Beers, P. J., Boshuizen, H. P. A., & Kirschner, P. A. (2007). The analysis of negotiation of common ground in CSCL. Learning and Instruction, 17, 427–435.

[3] Morgan, R. L., Whorton, J. E., & Gunsalus, C. (2000). A comparison of short-term and long-term retention: Lecture combined with discussion versus cooperative learning. Journal of Instructional Psychology, 27, 53–58.

[4] Van Boxtel, C. A. M., Van der Linden, J., & Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. Learning and Instruction, 10, 311–330.

[5] Kirschner, P. A., Martens, R. L., & Strijbos, J. W. (2004). CSCL in higher education? A framework for designing multiple collaborative environments. In P. Dillenbourg, J. W. Strijbos, P. A. Kirschner, & R. L.Martens (Eds.), Computer-supported collaborative learning: Vol 3. What we know about CSCL…and
implementing it in higher education (pp. 3–30). Boston, MA: Kluwer Academic.

[6] Heller, P.,Keith,R. & Anderson, S (1992). Teaching problem solving through cooperative grouping: Group versus individual problem solving. American Journal of Physics Teachers, 60, p 627 – 636.

[7] Sharan,Y., & Sharan,S. (1992). Expanding cooperative learning through group investigation. NY:Teachers College Press.

[8] Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (2nd ed.). Boston: Allyn & Bacon.

[9] Repman, J. (1993). Collaborative ,computer based learning: cognitive and affective domains, *Journal of Research on Computing in Education*, 9(2), p 149 – 163.

[10] Vygotsky, L.S (1978). *Mind and society:The development of higher psychological processes*. M.Cole, V.John Steiner, S. Scribner, & E Souberman (Eds & trans.)(Originally 1927) Cambridge; Harvard University press.

[11] Smith,K.A. (1992). *Cooperative/collaborative learning for active students involvement: Essentials of Cooperative Learning. Teaching Concerns*. URL: http://www.virginia.edu/~trc/tccoop.html).

[12]Johnson, D. W., & Johnson, R. T (1995). *Creative Controversy: Intellectual challenge in the classroom* (3rd edition). Edina, MN: Interaction Book Company.

[13] Kimber,.D. (1994) *Collaborative Learning in Management education: Issues and Benefits, problems and solutions: A Literature Review*.

[14]Springer, L., Stanne, M. E., and Donovan, S. (1998). "Effects of cooperative learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis." (Research Monograph No. 11). Madison: University of Wisconsin-Madison, National Institute for Science Education, *Review of Educational Research*.

Implementation of a Curriculum Management System at a Faculty of Medicine – Lessons Learned from a Pilot Project

Aarti Paul
Faculty of Medicine
University of Manitoba
paula@cc.umanitoba.ca

Daniel Yaworski
Faculty of Medicine
University of Manitoba
daniel_yaworski@hotmail.com

Keith McConnell
Faculty of Medicine
University of Manitoba
mcconnek@cc.umanitoba.ca

Bruce Martin
Faculty of Medicine
University of Manitoba
bmartin@cc.umanitoba.ca

Dean Sandham
Faculty of Medicine
University of Manitoba
Dean_Sandham@cc.umanitoba.ca

Sat Sharma
Faculty of Medicine
University of Manitoba
ssharma@sbgh.mb.ca

***Abstract*— Integration of electronic curricula remains highly variable among medical schools in North America. A major challenge has been availability and implementation of an ideal curriculum management software system (CMS). The objective of this study was to test the feasibility of using a CMS system at the University of Manitoba (U of M). At the U of M, we implemented a multi-user software system designed to provide faculty and students with an integrated curriculum, scheduling, calendar functionality, and communication tools. The goals of the pilot project were to explore the feasibility of an electronic curriculum and to evaluate strengths and weaknesses of the software. One block each from the 1st and the 2nd year medical school class were chosen for the pilot project. Data were collected through student and faculty surveys, student and faculty login-ins, student page views, and through direct communication with various users. Most students (100%) and faculty (78%) indicated a strong need for an electronic CMS. A statistically significant increase in student log-ins ($p<0.01$) and page views ($p<0.05$) occurred from week 1 to week 6 and then reached a plateau, whereas faculty log-ins remained constant. The 1st year and 2nd year medical classes were dissatisfied (34% and 57%), neutral (24% and 25%), or satisfied (42% and 18%) with the CMS, respectively. This pilot project demonstrated that an electronic CMS was feasible at the U of M. In addition, we learned how to effectively implement such a system, and how to encourage faculty and students to adopt the system. The lessons learned at U of M may be instructive for other faculties attempting to deploy electronic curriculum for undergraduate medical education.**

Keywords- eLearning; electronic curriculum; curriculum management system; learning management system.

I. INTRODUCTION

Medical education aspires to provide future doctors with the knowledge and skills to practice medicine and to develop lifelong learning skills [1], [2]. The traditional model of medical education is based on transferring scientific facts to students via lectures, notes and books, culminating in knowledge synthesis, and acquisition of skills and professional attitudes [3]−[5]. Recent advances in information technology have laid a strong foundation for the use of electronic technology to deliver, support and enhance teaching, learning, and evaluation [6]. Using well-designed computer-based courseware, electronic learning can facilitate live teaching with streaming lectures, whiteboards, downloadable slide sets, and discussion forums [6]−[9]. The benefits of e-learning include faster learning at reduced cost, increased access to learning, and accountability of all participants in the learning process [10]−[11]. Changes in health care delivery, advances in medicine, increased demands on academic faculty, and transformation of health care delivery sites from tertiary care institutions to community-based settings have challenged traditional teaching time and resources [12].

II. BACKGROUND WORK

Despite gaining popularity over the past decade, the use of e-learning remains highly variable among medical schools and its integration has been a genuine challenge, especially when teaching clinical disciplines [6]−[12]. Creation and deployment of learning management systems that support e-learning are arduous and expensive, and their integration into medical education is equally challenging. E-learning material involves several components [13]−[14]. Once the content is developed, it must be managed, delivered, and standardized. Content management includes all the administrative functions (e.g., storing, indexing, and cataloguing) needed to make e-learning content available to the learners [15]. Examples include portals, repositories, digital libraries, learning-management systems, search engines, schedules, and personalized e-portfolios. A curriculum management system includes web-based software that facilitates the delivery and tracking of e-learning across an institution. It may also serve several additional functions [16]−[18]. For example, CMS can simplify and automate administrative and supervisory tasks, track learner's achievement of competencies, and provide a

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_16,

readily accessible and life-long repository for learning resources [19]. The American Association of Medical Colleges has suggested various approaches to implement informatics skills into the medical school curriculum [1].

In 2004, the Faculty of Medicine (FOM) at the U of M began exploring the feasibility of acquiring a multi-user, integrated electronic curriculum management system for its medical school. The goal of this curriculum system was to provide medical students and faculty the best available technology for knowledge transfer. Such a system was expected to offer a simple, user-friendly interface that could be utilized by individuals with varying levels of technical skill. In this paper, we describe our experience and the challenges we encountered at U of M during selection, preparation, customization, implementation and evaluation of a web based comprehensive electronic curriculum management system.

III. Methods

A. Selection of the Curriculum Management System

We began by analyzing curriculum management systems currently in use at various Canadian and American medical schools. Several candidate software systems were carefully evaluated for functionality, architecture, support, and cost efficiency. Following a preliminary review, detailed and exhaustive analysis of three systems was performed. These included: 1) Curriculum Information System initially developed by the Med-IT Department, University of Calgary, and its intellectual property rights are currently owned by U of M.; 2) Web Eval, developed by One 45, a commercial medical informatics company; and 3) WebCT, commercial e-curriculum software.

The product analysis substantiated Curriculum Information System (CIS®) compared to Web Eval and Web CT is a comprehensive software package that links scheduling, learning objectives and resources. Successfully implemented previously at the University of Calgary, CIS® is an integrated electronic medical curriculum and scheduling software that can be delivered via the Internet and wireless hand-held PDA's. The goals of the pilot project were to explore the feasibility of implementing an electronic curriculum at U of M and to evaluate strengths and weaknesses of the CIS software. We envisioned a CMS that would foster and sustain lifelong learning for medical students by providing a solid undergraduate medical education foundation for future growth. An ideal system would be a multi-user, web based collection of electronic tools designed to assist medical faculty, students and administrators in the management of all knowledge pertinent to learning objectives defined by the curriculum.

B. Preparation for the Pilot Project Launch

Prior to the pilot project, we prepared and uploaded learning resources and teaching objectives to populate the system. All teachers were responsible for ensuring that their content was current, fulfilling the teaching objectives, and uploading content prior to the teaching activity. Most teachers uploaded their own material, while others e-mailed it to the course director, course administrator, or to us for uploading. Personal communication with the faculty and course directors appeared to be the most successful method of preparing the curriculum for the pilot project. An implementation committee was created to oversee the pilot roll-out. Members of the committee included directors from 1st and 2nd year medical school classes, course directors, pre-clerkship administrators, and student representatives from each class. Feedback and counsel was obtained from the committee and other users prior to customization of system for the U of M pilot project.
(Fig. 1)

We used multiple strategies to prepare medical school faculty, students, and administrative staff for a successful launch of the pilot project. These included personal communications, presentations, and discussion at various undergraduate administrative/education committees. Critiques and suggestions obtained from multi-user groups were collated, analyzed, and put into practice. Hands-on training sessions were scheduled and well attended by faculty, students, course directors, and their administrative staff. A videoconference was arranged to communicate, share and learn from the University of Calgary's experience with the implementation. They experienced numerous adoption difficulties as their curriculum gradually evolved into a paperless system. A focus group included 15 to 20 students from each class, CMS representatives, and academic representatives to enhance communication between the implementation team and the class.

After approximately six months of preparation and planning the CMS was launched in January of 2007. We targeted one block from each pre-clerkship year for the pilot project. Block III, taught during the 1st year consists of Respiratory Medicine, Cardiovascular and Otolaryngology courses. Block VI, launched in March of 2007 is taught during the 2nd year and includes Gastroenterology, Hematology, and Dermatology courses. There were 105 students in the 1st year class and 98 students in the 2nd year class. The launch consisted of fully integrated individualized schedules linked to the curriculum, discussion forums, news groups, calendar, and e-mail functionality. Since this was a pilot project, paper resources and student websites were maintained as before. Therefore, three simultaneous systems existed for the curriculum during the pilot project: paper-based notes; teaching material posted on the student website; and CMS. Two systems coexisted for scheduling: a printed course schedule book and CMS.

A pilot project evaluation committee was created to evaluate the feasibility of an electronic curriculum information system for the FOM, U of M, and to assess strengths and weaknesses of the CMS software. The committee contained representatives from 1st and 2nd year classes, academic and CMS representatives, 1st and 2nd class year program coordinators, course directors for Block III and VI, an education expert, and administrative coordinators. Feedback was solicited from the students, faculty, course directors and administrative staff through student and faculty surveys,

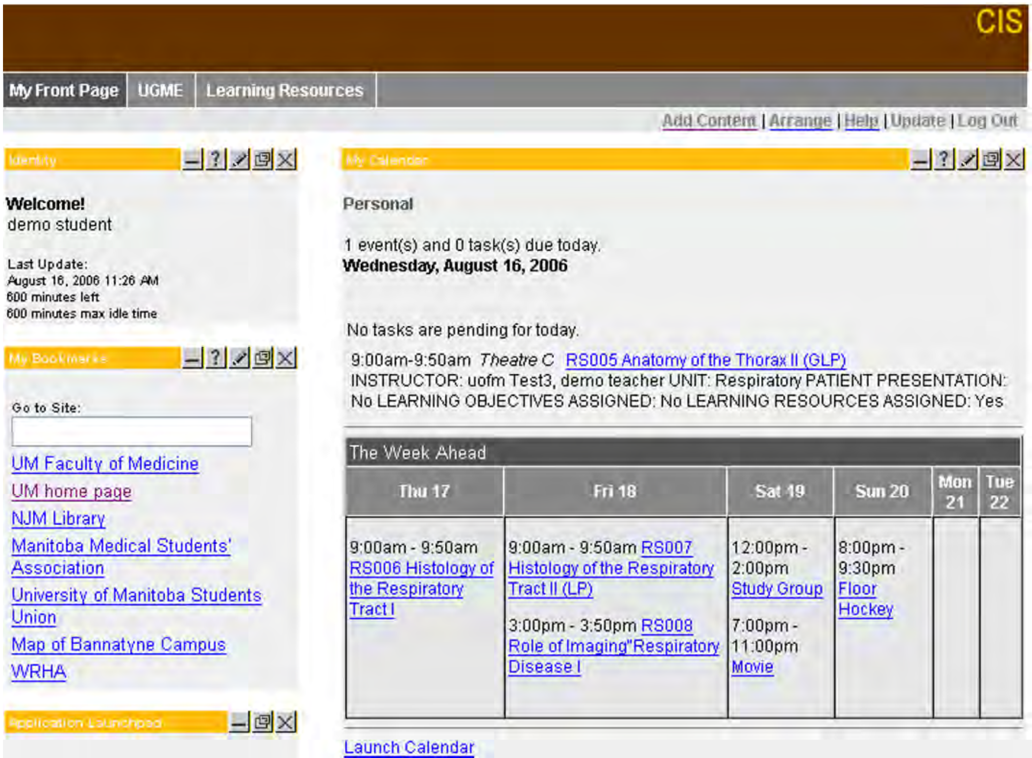


Fig.1. A personalized interface for students and faculty opens up as the user logs in.

student and faculty log-in and page view data, and informally through direct communication with the users.

We hypothesized that a substantial increase in log-ins and page views over time will demonstrate success, especially since multiple systems were available. Data from week 1 was compared to data from week 6. The sequential log-in and page view data was analyzed using chi-square analysis and student's t-test.

IV. Results

Weekly system usage data were extracted from the system logs and included a breakdown of various components and modules. While faculty logins remained constant, student logins continued to increase for both 1st and 2nd year classes. A statistically significant increase ($p < 0.01$) in student logins occurred from week 1 to week 6 and then reached a plateau **(Fig. 2).** Each student logged into the system approximately once a day, including the weekends. This data is remarkable since login was not necessary or required for the courses. Each student received paper and other resources, including the scheduling information and course material in advance. Furthermore, with one login, students could access and download all teaching material for the day. Curriculum page views indicate that a learning resource was accessed by the user. These page views dramatically increased over time for both 1st and 2nd year classes. Major increases ($p < 0.05$) occurred between weeks 1 and 6, and the access remained stable thereafter. **(Fig. 3)** An independent survey conducted by the students was submitted to the evaluation committee. The survey evaluated student satisfaction, and strengths and weaknesses of the CIS® software.

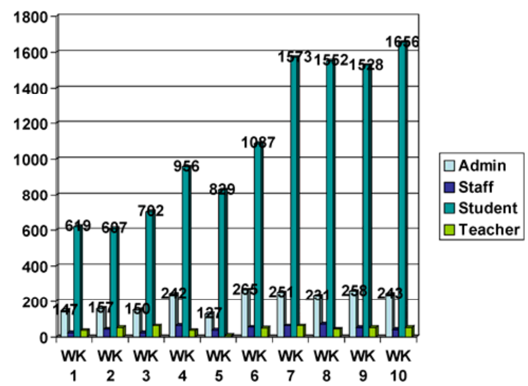


Fig. 2. Login data for 1st year class from week (WK) 1 to week 6 and both 1st and 2nd year classes from week 7 to week 10. A significant increase in log-ins ($p < 0.01$) occurred from week 1 to week 6.

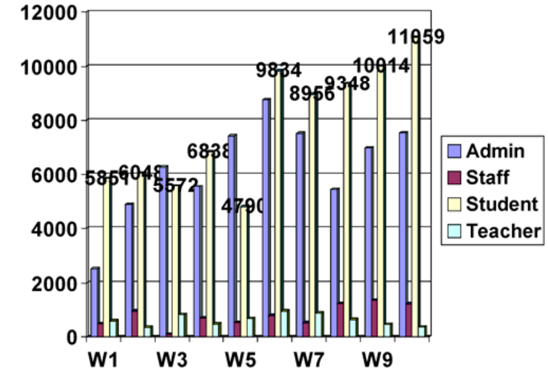


Fig. 3. Curriculum page views for 1st year class from week (W) 1 to week 6 and both 1st and 2nd year classes from week 7 to week 10. A significant increase in page views (p <0.05) occurred from week 1 to week 6 and week 10.

A. Analyzing Students' Responses

It is important to analyze student evaluations and comments in an appropriate context. The student acceptance and adoption that we achieved during this pilot project met our expectations. The project was conducted in the middle of an academic year for the 1st year students after they had adjusted to the paper based scheduling and curriculum system. The 2nd year students were generally less enthusiastic as this was their last pre-clerkship block before starting clinical rotations. Students also had ready access to the class website, an unmonitored site that serves as a resource for communication and planning social events. The class website is also used as a repository for learning materials. The student website is much easier to navigate than the CMS which requires training and has a learning curve to become proficient. While most students were either supportive or neutral **(Table 1)**, some were resistant to change and thus were the most vocal opponents of the pilot project. Therefore, "cultural issues" or inherent "skepticism" for change definitely appeared to be an important factor in adoption. Such barriers during the implementation of electronic systems are well documented in the literature and gradually decline before widespread acceptance and adoption [20]–[21].

Feedback from the Faculty

A total of 53 instructors from 6 clinical disciplines participated in this pilot project. Six of the course directors were intimately involved with the roll-out and evaluation. All course directors and many faculty members provided formal and informal feedback on the electronic curriculum and evaluation of the software. Faculty response was generally positive and enthusiastic. Many course instructors used the system extensively for scheduling functions and to update and enhance teaching resources. Results of a survey sent to all faculty members involved with the pilot project at the end of the pilot project are shown. **(Fig. 4 and 5)**

TABLE 1

Results of an independent survey conducted by the 1st and 2nd year students. This table shows data from student's responses to the question "How would you rate your overall satisfaction with the CIS?"

Class	Very Dissatisfied	Somewhat Dissatisfied	Neutral	Somewhat Satisfied	Very Satisfied
1st Year 76 (72%) Responses	11%	23%	24%	29%	13%
2nd Year 79 (81%) Responses	32%	25%	25%	16%	2%

V. Recommendations from the Pilot Project

Based on analysis of the pilot data and comprehensive consultation with all stakeholders, the evaluation committee made several recommendations. These lessons learned at U of M may assist other medical schools in their own development and implementation of electronic curricula.

1) To prepare future physicians to deliver best medical care, the faculty and students at U of M determined that an electronic curriculum and an integrated curriculum management system were high priorities.
2) The pilot project at U of M provided a unique opportunity for students, faculty, and administrative staff to conceptualize, prepare, launch, troubleshoot, and evaluate an electronic curriculum. Extensive preparation, free communication, thorough consultation and engagement of all users from an early stage and throughout are required for successful implementation.
3) The pilot project at U of M identified weaknesses of the software system and gaps in the wireless infrastructure and hardware of the FOM. **(Table 2)** This evaluation guided further modifications and an upgrade of the CIS® software to improve interface, usability, functionality and speed of the system.
4) To gain faculty-wide engagement and adoption, duplication of the curriculum and schedule should be

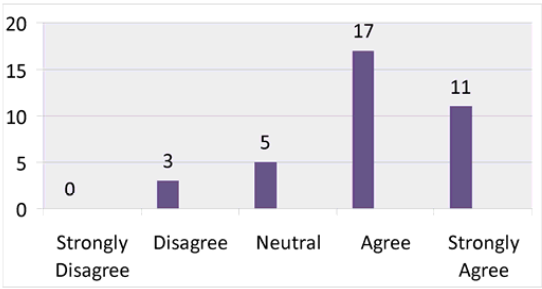


Fig. 4. Faculty response to the question: "I believe in the concept of an Electronic Curriculum for the Faculty of Medicine at U of M"

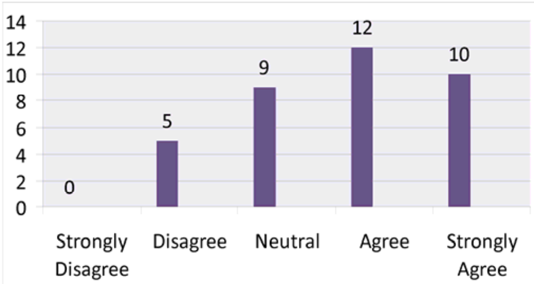


Fig. 5. Faculty response to the question: "I find it easy to log-on, navigate and upload content on CIS"

avoided and mandatory hands-on training sessions will likely enhance the user experience and overall success of the electronic curriculum.

VI. Discussion

Many medical schools in North America and Europe have published their experience with deployment of electronic curricula for either a single course or the whole curriculum. We reviewed their experience and related it to the lessons that we learned during the pilot project. Our experience showed that gradual, controlled and systematic (Change Management) rather than outright adoption is preferred as users develop proficiency and familiarity with the on-line system.

The University of British Columbia Faculties of Medicine and Dentistry have developed a MEDICOL (Medicine and Dentistry Integrated Curriculum Online) site for medical students to use during their rural family practice, a four- to six-week experience in the summer after the second year [22]. The system uses WebCT, tracks student progress, promotes interactions through e-mail and bulletin boards, and delivers self-directed learning components and multimedia learning modules. The use of statistics indicates that over 90% of students regularly used the MEDICOL sites and have found them helpful.

A case study to "electrify the curriculum" at Virginia Commonwealth University School of Medicine chose a second-year course to serve as a model e-course and to provide evaluation data for a two-year study [23]. E-mail surveys, focus groups, student diaries, and comprehensive end-of-course student assessments showed that students found access to multiple images, interactivity, and meaningful and efficient navigation within the site to be useful tools to enhance complex learning. Gotthardt et al. recently described a strategy of successfully implementing an e-learning curriculum in the Netherlands [24]. The curriculum consisted of self-directed learning, an online discussion forum, and discussion rounds. The e-content in nuclear medicine and radiotherapy was produced by a team of medical authors, web designers, and psychologists, and was delivered via an online learning management system. Online evaluation proved that the new curriculum was very successful, as the efficacy of teaching increased and the duration of the course could be reduced by half. These publications present compelling evidence that electronic curricula and systems are viable and can be successfully used to deliver medical education.

TABLE 2

Results of an independent survey conducted by the 1st and 2nd year students. The strengths and weaknesses of the electronic curriculum system CIS) are listed here. The data is combined for 1st and 2nd year class.

% of Responses	Strengths
100%	Electronic curriculum system is a very good idea
100%	University of Manitoba absolutely needs a curriculum management system
14%	CIS has good functionality
87%	CIS is a multi-user system with well linked scheduling and learning resources
92%	Personalized scheduling is a big plus for the CIS system
% of Responses	**Weakness**
46%	CIS interface is not very attractive and user friendly
69%	System is slow at times and requires multiple mouse clicking
62%	Extra windows need to be opened in the system to access learning resources
74%	Calendar view and leaning resources cannot be linked
59%	System is not intuitive and requires training to be proficient

In spite of the enthusiasm and great potential for e-curriculum and curriculum management systems, insufficient literature exists to assist medical schools with implementation strategies. Authors of one study evaluated mechanisms that may assist with implementation of an e-curriculum to enhance dental education [25]. Electronic Curriculum Implementation Survey (ECIS), a 26 item questionnaire was mailed to all sixty-six North American dental schools (ten Canadian and fifty-six U.S. schools) during 2002-03 with a response rate of 100 percent. The survey found that e-curriculum implementation among North American dental schools followed the classic innovation pattern. A few early adopting institutions proceed rapidly while the majority of potential adopters make modifications slowly. The implementation barriers identified by the survey were: lack of time, skill, and incentive to develop the educational software. Another educational study by Riley at al. evaluated the effectiveness of the web-based teaching model as applied to perfusion education [26]. Web-based virtual classroom (VC) environments and educational management system (EMS) software were implemented independently and in addition to the live, interactive Internet-based audio and video transmission. Use of Internet-based VC and EMS software was a significant improvement over traditional models. Interestingly, students recognized the learning efficiency of on-line information but still preferred the traditional face-to-face classroom environment. A Spanish study delivered and evaluated an educational web based cardiovascular course prepared as adjunct educational material for the medical student class [27]. Frequency of use, impact on in-class attendance and student satisfaction, number of visits and quality of the website were assessed. The quality score rating was a mean of 7.7 (scale: 1-10); 93.4% students attended the class, and 88.2% of the students evaluated the web as a useful or very useful addition to medical teaching.

Kerfoot et al. conducted a multi-institutional randomized controlled trial to investigate the impact of an adjuvant web-based teaching program on medical student learning during clinical rotations [28]. Web-based teaching cases were developed covering four core urologic topics. Students were block-randomized to receive web-based teaching on two of the four topics. Before and after a designated duration at each institution, students completed a validated 28-item web-based test covering all four topics. Compared to controls, web-based teaching significantly increased test scores in the four topics at each medical school ($p < 0.001$, mixed analysis of variance), corresponding to an effect size of 1.52 (95% confidence interval [CI], 1.23-1.80). Learning efficiency increased three-fold by web-based teaching (effect size 1.16; 95% CI 1.13-1.19). Furthermore, students who were tested a median of 4.8 months later and those who received web-based teaching demonstrated significantly higher scores than those who received non-web-based teaching ($p = .007$, paired t-test). This randomized controlled trial provides convincing evidence that web-based teaching can significantly and durably improve medical student learning.

Our experience developing and implementing an e-curriculum at U of M is similar to the journey undertaken by McGill University [2]−[29]. David Fleiszer and Nancy Posel were the leaders in development of an online, multimedia-enhanced undergraduate medical curriculum. The electronic curriculum aimed to include: (1) learner-centered, self-directed, and guided learning; (2) inherent interactivity of the technology; (3) a powerful means for integrated learning; (4) ability to address a variety of learning styles experienced by the students; and (5) opportunities for horizontal and vertical curricular integration. The experience at McGill is ongoing and has been successful partly because of a substantial private funding grant and strong leadership.

VII. Conclusion

The vision of electronic learning is to foster lifelong learning for future physicians beginning with robust undergraduate medical education that emphasizes strong self-learning skills and academic scholarship [30]−[32]. E-learning has been shown to be efficient, effective and has received acceptance within the medical education community. Its success is only possible via an integrated, user-friendly, versatile and highly functional CMS. Such systems are currently in development and not widely available. Considerable effort and resources are required for individual institutions to customize and implement a CMS. Our experience with a pilot project using an electronic curriculum management system (CIS®) established a significant need for such a system, demonstrated that implementation is feasible, and found that adoption by faculty and students can be expected. Successful implementation of e-learning in medical education mandates a thorough consultation, effective communication and engagement of all users.

VIII. Acknowledgment

We would like to acknowledge and thank class of 2011 students, Blocks III and VI instructors and administrative staff of Undergraduate Medical Education Office (Pat McCullough, Pre-Clerkship Program Administrator; Mark Torchia, Med I Coordinator; and Vieanna Beck, Administrative Assistant) at the Faculty of Medicine, University of Manitoba for their superb enthusiasm and terrific support.

References:

[1] Effective use of educational technology in medical education. AAMC 2007. https://aamc.org/ Accessed on August 8, 2008.

[2] D. Sandham, A. Paul, S. Sharma. Integration of e-curriculum into medical education. Gravitas 2008; 41:4-5.

[3] Educating doctors to provide high quality medical care - a vision for medical education in the United States. AAMC 2004. https://aamc.org/ Accessed on August 8, 2008.

[4] B. Barzansky, S.I. Etzel. Educational programs in US medical schools, 2002-2003.

[5] H.B. Slotnick. How doctors learn: education and leaning across the medical-school-to- practice trajectory. Acad med 2001; 76: 1013-26.

[6] J.P. Ward, J. Gordon, M.J. Field, H.P. Lehmann. Communication and information technology in medical education. The Lancet. 2001; 357:792–796.

[7] P. Devitt, E. Palmer. Computer-aided learning: an overvalued educational resource? Med Edu 1999; 33:136-139.

[8] T. Greenhalgh. Computer-assisted learning in undergraduate medical education. BMJ. 2001; 322:40–45.

[9] C.L. Friend, C.L. Cole. Learner control in computer-based instruction: a current literature review. Ed Tech. 1990; November: 47–40.

[10] D.J. Steele, J.E.J. Palensky, T.G. Lynch, N.L. Lacy, S.W. Duffy. Learning preferences, computer attitudes, and student evaluation of computerized instruction. Med Educ. 2002; 36:225–232.

[11] C.J. Bland, S. Starnaman, L.Wersal, L. Moorhead-Rosenberg, et al. Curricular change in medical schools: how to succeed. Acad Med 2000; 75:575–94.

[12] M. Vogel, D.F. Wood. Love it or hate it? Medical students' attitudes to computer-aided learning. Med Educ. 2002; 36:214–215.

[13] M.L. Rice, E.K. Wilson, W. Bagley. Transforming learning with technology: lessons from the field. J Tech Teacher Educ 2001; 9(2):211–30.

[14] B.M. Karnath, W. Thornton, M. DasCarlos. Pilot study of computer-based self-teaching in cardiac auscultation. Med Educ. 2003; 37:1048–1049.

[15] A.J. Lipman, R.M. Sade, A.C. Glotzbach, C.J. Lancaster, M.F. Marshell. The incremental value of internet-based instruction as an adjunct to clasroom instruction: a propsective randomized study. Acad Med. 2001; 76:1060–1064.

[16] B.L. Ingerman. Form follows function: establishing the necessary infrastructure. In: Barone CA, Hagner PR, eds. Technology-enhanced teaching and learning. San Francisco: Jossey-Bass, 2001.

[17] J. Daniel. Computer-Aided Instruction on the World Wide Web: The Third Generation. J of Econ Ed. 1999. pp. 163–174.

[18] S. Chodorow. Educators must take the electronic revolution seriously. Acad Med. 1996; 71:221–226.

[19] H.K. Muhsin. The use of e-learning techniques to improve the medical education. ICTTA 2008. http://ieexplore.ieee.org/ Accessed on August 2008.

[20] R.M. Palloff, K. Pratt. Lessons from the cyberspace classroom: the realities of online teaching. San Francisco: Jossey-Bass, 2001.

[21] H.S. Wentzer. Technology development and implementation in the public health institutions and pedagogical task for the learning organization. Stud Health Tech Informatics 2000; 77:264–70.

[22] M. Broudo, C. Walsh. MEDICOL: online learning in medicine and dentistry. Acad Med. 2002 Sep;77(9):926-7.

[23] J.D. Hoban, J.B. Schlesinger, R.P. Fairman, M.M. Grimes. Electrifying a medical school course: a case study. Teach Learn Med. 2003 Spring;15(2):140-6.

[24] M. Gotthardt, M.J. Siegert, A. Schlieck, S. Schneider, A. Kohnert, M.W. Gross, et al. How to successfully implement e-learning for both students and teachers. Acad Radiol. 2006 Mar; 13(3):379-90.

[25] W.D. Hendricson, F. Panagakos, E. Eisenberg, J. McDonald, G. Guest, P. Jones, et al. Electronic curriculum implementation at North American dental schools. J Dent Educ. 2004 Oct; 68(10):1041-57.

[26] J.B. Riley, J.W. Austin, D.W. Holt, B.E. Searles, E.M. Darling. Internet-based virtual classroom and educational management software enhance students' didactic and clinical experiences in perfusion education programs. J Extra Corpor Technol. 2004 Sep; 36(3):235-9.

[27] X. Gómez-Arbonés, A. Ferreira, M. Piqué, J. Roca, J. Tomás, J.L. Frutos, et al. A cardiological web as an adjunct to medical teaching: prospective analysis. Medical Teacher 2004; 26:187-191.

[28] B.P. Kerfoot, H. Baker, T.L. Jackson, W.C. Hulbert, D.D. Federman, R.D. Oates, et al. A multi-institutional randomized controlled trial of adjuvant web-based teaching to medical students. Acad Med. 2006 Mar; 81(3):224-30.

[29] D.M. Fleiszer, N.H. Posel. Development of an undergraduate medical curriculum: the McGill experience. Acad Med. 2003 Mar; 78(3):265-9.

[30] S. Schultze-Mosgau, T. Zielinski, J. Lochner. Web-based, virtual course units as a didactic concept for medical teaching. Med Teach. 2004 Jun; 26(4):336-42.

[31] J.G. Ruiz, M.J. Mintzer, R.M. Leipzig. The impact of e-learning in medical education. Acad Med 2006: 81:207-212.

[32] K. Masters, T. Gibbs. The spiral curriculum: implications for online learning. BMC Med Educ. 2007 Dec 21; 7:52-55.

Use of Ontology for Reusing Web Repositories for eLearning

Punam Bedi, Hema Banati, Anjali Thukral

***Abstract*— eLearning scenario has witnessed a shift from instructional to adaptive learning. The focus is now moving from closed to open scenario. However, learning in such situation relegates to the task of retrieving as the information is largely scattered across the web. Searching such voluminous information is practically infeasible and leads to incomplete or irrelevant results. Hence, a need is felt to automatically integrate the vast pool of information present in varying forms on web. This paper proposes a framework that searches for information semantically across all existing web repositories for a given query. The approach utilizes WSD (word sense disambiguation) algorithm to semantically analyze and expand the query. A personalized knowledge base for learning purpose is subsequently created by deploying agents to compile information from various learning resources. The integrated organization of knowledge provides a single point reference for the posted query. It thus increases the probability of the learner obtaining complete and relevant information from a single reference. This framework therefore allows a learner to concentrate on learning than the searching process thus enhancing the learning experience.**

Keywords-eLearning; eLearning; web repositories; query expansion; software agents; personalized ontology.

I. Introduction

In the past practice traditional systems like LMS, LCMS [1, 2] and other intelligent systems like AWBES [3] have played an important role in the evolution of eLearning [4]. These systems emphasis on providing an integrated environment for a learner, but within a separate and closed scenario, viz within an educational organization or within some paid association. This restriction significantly narrows the educational inputs, depriving a learner from vast pool of information thus precluding from the efficiency of learning [5]. Learning in open scenario is also long-winded. The presence of voluminous and scattered information in varying formats deviates and hinders the learning process of a learner. A need is hence felt for a system that can automatically and dynamically organize the information from all possible resources to a well formed structure. This structured learning material helps a learner for a better understanding of concept. This paper, therefore addresses above mentioned need and proposes an eLearning framework that provides latest updated learning material, semantically searched from web repositories for the entered query.

Manuscript submission date: 26th October 2009.
Punam Bedi,(pbedi@cs.du.ac.in), Hema Banati,(hema.banati@gmail.com)
Anjali Thukral (anjali.thukral@yahoo.co.in)
Dept. of Computer Science, *University of Delhi, India*

The paper is organized as follows: Section II presents the background work, briefly defining eLearning, advantages of Semantic Web (SW) for it and ontology. Section III describes the framework and algorithm for reusing web repositories for eLearning. Finally section IV discuss future work and concludes the paper.

II. Background Work

eLearning is defined with eight dimensional characteristics as: *Pull*: Learner determines agenda (delivery), *Reactionary*: Responds to problem at hand (responsiveness), *Non-linear*: Allows direct access to knowledge in whatever sequence makes sense to the situation at hand (access), *Symmetric*: Learning occurs as an integrated activity (symmetry), *Continuous*: Learning runs in parallel to business tasks and never stops (modality), *Distributed*: Content comes from the interaction of the participants and the educators (authority), *Personalized*: Content is determined by the individual user's needs and aims to satisfy the needs of every user (personalization) and *Dynamic*: Content changes constantly through user input, experiences, new practices, business rules and heuristics (adaptivity) [6,7]). eLearning characteristics can effectively be achieved through the new WWW architecture, Semantic Web [8] (SW). Research [6, 7, 9, 10, 11, and 12] shows the benefits of using semantic web as the platform for eLearning. Conceptually the SW represents a layered architecture (Fig 1), where each level provides different degree of expressiveness. The bottom most layer allow user to provide a controlled vocabulary and namespaces. Next layer, Resource Description Framework (RDF) represents information about resources in the World Wide Web. It consists of a triplet – resource, property and value. Ontology vocabulary, the next upper layer, adds more semantics to RDF by allowing a better specification of constraints on concepts. The rules for explicit inference are established by Logic layer which are then executed by the Proof layer. The top most layer, Trust provide mechanisms to imbibe trust in a given proof [13]. The semantic web as whole, acts as a decentralized knowledge bank that provides meaning to a learner's search and thus allows extracting most appropriate learning material.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_17, © Springer Science+Business Media B.V. 2010

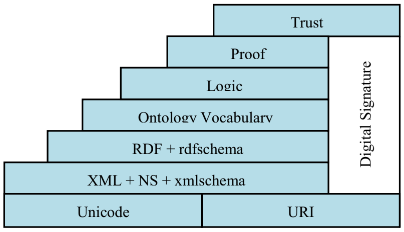


Figure 1: Layered Architecture of Semantic Web

The middle layer and the backbone of Semantic web, Ontology Layer, forms knowledge about a certain domain by providing relevant concepts and relations between them. Often cited definition of ontology is "an explicit specification of a conceptualization" [14] i.e., an ontology defines (specifies) the concepts, relationships, and other distinctions that are relevant for modeling a domain. The specification takes a form of representational vocabulary (classes, relations, and so forth) which provides meaning for the vocabulary and formal constraints on its coherent use [15].

Mathematically, a core Ontology is defined as a structure, [16]

$$O := (C, \leq_C, R, \sigma, \leq_R)$$

consisting of:

i) two disjoint sets C and R whose elements are called concept identifiers and relation identifiers, respectively.
ii) a partial order $\leq_C$ on C, called concept hierarchy or taxonomy,
iii) a function $\sigma : R \rightarrow C^+$ called signature, and
iv) a partial order $\leq_R$ on R, called relation hierarchy, where $r_1 \leq_R r_2$ implies $|\sigma(r_1)| = |\sigma(r_2)|$ and $\pi_i(\sigma(r_1)) \leq_C \pi_i(\sigma(r_2))$ for each $1 \leq i \leq |\sigma(r_1)|$

and Knowledge Base is defined as a structure [16]:

$$KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R)$$

consisting of:

i) two sets C_{KB} and R_{KB} ,
ii) a set I whose elements are called instance identifiers (or instances or objects for short),
iii) a function $\iota_C : C_{KB} \rightarrow \beta(I)$ called concept instantiation,
iv) a function $\iota_R : R_{KB} \rightarrow \beta(I^+)$ called relation instantiation.

Ontologies have the potential to construct a dynamic and semantic structure of e-content that can be shared, reused and enhanced with trust for learning purposes. Since the invention of semantic web in 2000, many researchers have proposed various architectures and frameworks that use this technology to enhance the learning experience of learners [17]. Most of these frameworks initially started with the use of metadata. Efforts were made to convert eLearning courses to metadata on a large scale [18] to facilitate a vast group of learners. Standards were established to generalize the structure for such metadata as Dublin Core [19] and LOM (Learning Object Metadata) [20]. However, there were issues in using LOM. They represent static repositories and suffer from various shortcomings [21]. The ontology based metadata provided a perfect blend of eLearning systems and ITS (Intelligent Tutoring Systems) [22] in closed scenario, overcoming most of the earlier shortcomings. Subsequently, the ontology layer was identified as a standard layer of the knowledge based systems [23]. Existing work of open learning systems either provides links to the World Wide Web documents, based on refined query searches or exclusively build on semantic web. Semantic web conveys meaning to the content whereas existing WWW has evolved as a powerful information portal.

III. FRAMEWORK

This paper proposes an open learning framework that provides the leverage of both worlds. Unlike other existing systems, it dynamically integrates all available resources on web to construct ontologies. This integration is established on the posted query. Words contained in the query are semantically analyzed for sense based query expansion. Expanding query in this way improves the probability of retrieving relevant information drastically [24]. Thus, the system first formulates the query and then integrates resources to form an exhaustive and updated ontology.

This framework is based on three types of ontologies:

- NLO (Natural Language Ontology): It contains lexical relations between the language concepts. They are large in size and do not require frequent updates. They represent the background knowledge of the system to understand the query formed in the natural language. The framework uses NLO, WordNet to expand the original query.
- DO (Domain Ontology): It captures knowledge of one particular domain, e.g., operating system, computer etc. These ontologies provide detailed description of the domain concepts. In this framework it belongs to the horizontal upper layer of the two level ontology [25] structures. (Fig 2).
- OI (Ontology Instances): They are the main resource of knowledge presented in the web and on the learner's local machine. These instances consist of interlinked web resources (includes web pages etc.) forming a horizontal hierarchy (lower layer of two level organization - Fig 2) among themselves and also vertically linked to their related concepts in the domain ontology. These are generated automatically and updated frequently by software agents. DO and OI forms the local knowledge base.

The framework (illustrated in Fig: 2) initializes the process by expanding query through query engine. The expanded query is then used to extract contents from web repositories. This retrieved information is ontologically organized to form

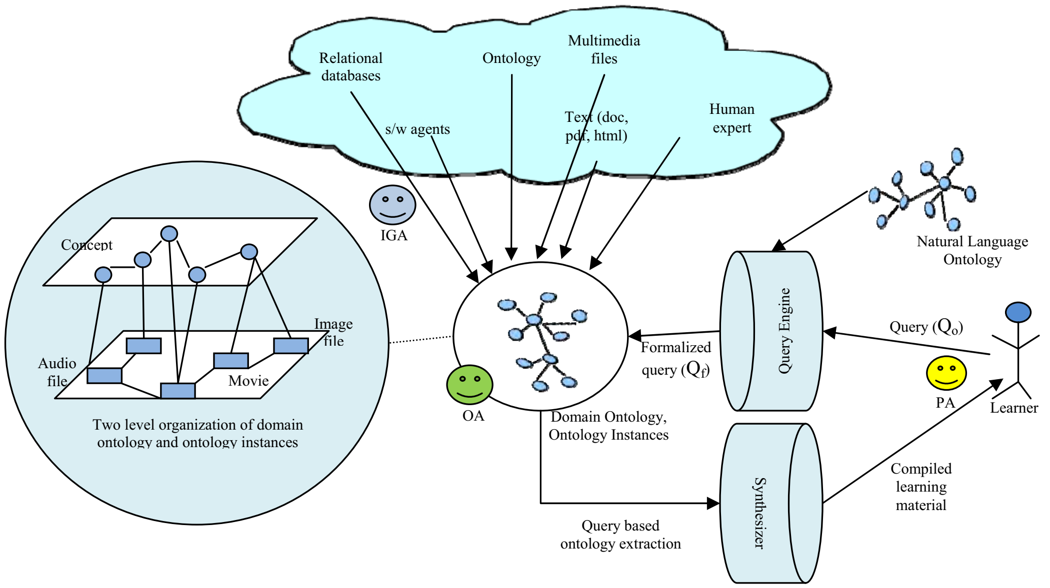


Figure 2: Framework- reusing web repositories for eLearning using ontology

knowledge base. The output is subsequently re-framed to a presentational format by the Synthesizer.
Remaining section explains the process of query expansion and content extraction followed by algorithmic summarization of the complete process of information retrieval.

A. Query expansion

The query engine (ref Fig: 2) lexically extract words from the query for the expansion. As an example, let the query be: "medium used in Ethernet". It generates a set of words from original query as,

$$Q_o = \{\text{medium, Ethernet}\}$$

In general,

$$Q_o = \{ w_1 , w_2 ,..., w_k \}$$

where, w_k are the concepts extracted from the entered query. The query engine formalizes the query using a mapping function, ψ. The function uses WordNet[1] and Word Sense Disambiguation (WSD) algorithm [24] to expand the query. The expanded query consists of semantically analyzed set of words that results from the maximum number of intersecting nodes in the semantic networks, generated from WordNet synonym sets (synsets) of w_k , k = 1, .. , n. Thus, resulting in a formulated query,

$$Q_f = \{ w_1 , w_2 ,..., w_k , C_x \}$$

1 http://wordnet.princeton.edu/

where C_x is the best possible configuration of senses for Q_o . One of the possible expansions for the above query can be seen as:

Q_f = {medium, Ethernet, coaxial cable, radio frequency signal, wireless local area network}.

The words, medium and Ethernet have been semantically analyzed to deduce new words that are used as medium for Ethernet. Though, the words like co-axial cable etc. does not exist in the original query, still the learner benefits by obtaining the intended information related to these supplementary words.

B. Content extraction from web repositories

Web repositories contain learning material in varying formats. IGA (Information Gathering Agent) manages the searching process to gather information from all possible resources. This information is organized semantically by OA (Ontology Agent) to update local personalized knowledge base. OA deals with each source format using specific procedure to give information an ontological structure. Following are the different formats handled by OA (ref. Fig. 2).

i) *Text documents*: The automatic generation of ontology from text (ontology population) [26] is achieved with KAON (Karlsruhe ontology) based tool TextToOnto [16]. The agent inputs text as a corpus and generates ontology using this tool. Generated ontology is later pruned or enhanced through the domain expert, if needed. The automatically generated ontology does not produce satisfactory results, but nevertheless provide some basic structure of concept hierarchy. Generation of quality ontologies without the human intervention is still a topic of research.

ii) *Multimedia files*: Ontology instances are used to organize multimedia files (audio files, video files, presentations, images etc.). These files are arranged in ontological form and at the same time linked to its related concept in Domain Ontology (DO).

iii) *Human experts*: Human experts' knowledge can be communicated through e-mail, blogs of FAQs (Frequently Asked Questions), or chatting. This is achieved by linking useful urls in the ontology instances and then linking them to related concepts in DO. Software agents [27] can be trained to take care of such conversations on behalf of the passive learner.

iv) *Ontology:* Reusing existing ontologies offer a much cheaper alternative than building new ones from scratch [28]. It provides one of the valuable repositories on web and reduces efforts and time to retrieve relevant content. However, the problem of determining relevance of the ontologies is not much different from finding a relevant searched document. A brilliant approach for automatic ontology construction from the existing ontologies on the web was proposed [28]. It extracts the relevant parts of the top ranked (ranked according to some criteria) ontologies and merge those parts to acquire the richest domain ontology. The ontologies on the web are searched using the semantic web search engines like Swoogle or Ontosearch. This approach requires further exploration of areas such as ontology ranking, segmentation, mapping, merging and evaluation.

v) *Software Agents*: Software agents are used by a system for personalized search and interaction with almost each form of web repository. As the future vision, one may find s/w agents on the web, similar to the human learner. It will be a normal scenario to have web full of s/w agents communicating to each other and negotiating for accessing relevant content for its owner. The challenge is to form a force of such good behaviored servants on web!

vi) *Relational databases*: Web repository also exists in the form of database. Majority of data on old systems consist of databases. Such resources of eLearning cannot be neglected. This framework maps the relational schemata of eLearning databases found on web, to ontology according to their underlying intrinsic relationships [29]. Tables, Fields and records are mapped to the ontology as class (concept), property and instances. They are further adjusted to get accommodated in the structure of proposed ontology.

All of the retrieved information from web repositories by above methods is organized in two level ontological structures to form a learner's personalized knowledge base. This knowledge base is then used further to answer queries related to same concept. The following algorithm generates a personalized knowledge base for a given query.

i) A learner enters a query in natural language, Q_o .

ii) A learner's personalized interface agent (PA) explores query entered by learner to infer the context and semantic.
   a) Through query engine, Q_o query is formalized by PA.
   b) The query engine extracts most appropriate words by applying Natural Language Ontology (WordNet) and WSD algorithm [24]. Thus a mapping, $\psi : Q_o \rightarrow Q_f$ is formed where, ψ is a mapping from the original query, Q_o to a formalized query, Q_f consisting of a set of semantically expanded words (explained in section III/A).
   c) PA then, submits Q_f , a formalized query to Ontology Agent, OA.

iii) OA performs two tasks simultaneously:
   a) First, it searches the most relevant ontology in its local knowledge base consisting of domain ontologies and ontology instances.
   b) Secondly, it instructs Information Gathering Agent, IGA to explore learning material in the World Wide Web and semantic web using Q_f (explained in section III/B).
   c) IGA returns to OA with:
      1) a set of top n ranked links (URLs), L_n after searching WWW and
      2) top m ranked ontologies, O_m after searching semantic web.

iv) Steps i) and ii) are repeated by PA till the learner is active or wishes to update its knowledge base passively. Consequently, PA keeps the learner updated by communicating with OA, and OA subsequently keep its knowledge base updated by frequently communicating with IGA.

v) OA update its knowledge base, KB with L_n and O_m to finally generate the result.

vi) Result is provided to the learner in one or both of the two forms:
   a) a set of ontologies ranked according to the relevance,
   b) top ranked ontology, forwarded to the 'Synthesizer', in order to generate a readable document with all its instances as links for the learner.

The above algorithm represents a complete process of reusing information by extracting only relevant content and providing single point reference to the compiled information for the learner. By using this algorithm, the framework creates an open scenario for learning, where a learner automatically gets the relevant and dynamically updated information related to the required concept without getting deviated from the flow of learning. This allows a learner to concentrate more on the concept understanding instead searching for the relevant content, which effectively enhances the learning experience.

IV. CONCLUSION AND FUTURE WORK

A framework has been proposed that reduces the chances of loosing out relevant content available on web significantly to create a personalized ontology for learning purpose with the additional vision of agent communication. It takes a query from learner, semantically expands it and gathers content from web repositories with the assistance of multiple software agents. These agents work for the active learner as well as searches content for the passive learner. Moreover, a personalized agent is provided to assist the learner at various learning points. At present, the framework has been partially implemented that uses JADE and a tool TextToOnto for extracting text and transforming it to ontology automatically. In future, other modules will be implemented and explored further to provide a more relevant ontology generation for the learner's knowledge base. A more formalized agent communication on the web and setting up a descent behavior will also be considered.

REFERENCES:

[1] S.R. Robbins, "The evolution of the Learning Content Management System", ASTD's Source for E-learning, 2002.

[2] Brandon Hall Research Article, 2007 [Online]: hhttp://www.brandon-hall.com/

[3] Peter Brusilovsky, "KnowledgeTree: A Distributed Architecture for Adaptive E-Learning", ACM 1-58113-912-8/04/0005, May 17-22, New York, USA.

[4] Anjali Thukral, Richa Sharma, Hema Banati, Punam Bedi, , "Modeling Roles in E-learning: MoRel", ACVIT'07 - International Conference on Advances in Computer Vision and Information Technology,612-620, Aurangabad, November 28-30, 2007.

[5] Michael G, Dimitry V.,Pavel, Denis, "Intelligent Educational Environments Based on the Semantic Web Technologies", Proceeding, IEEE International Conference on Artificial Intelligent Systems (ICAI'02) 0-7695-1733-1/02, 2002.

[6] Drucker, P. ,"Need to Know - Integrating e-Learning with High Velocity Value Chains". Delphi Group White Paper. 2000 www.delphigroup.com

[7] Yas A. Alsultanny, "eLearning System Overview Based on Semantic Web", The Electronic journal of e-Learning, Volume 4 Issue 2 (111-118) 2000.

[8] T. Berners-Lee, J. Hendler, O. Lassila , "The Semantic Web", Scientific American, May 2001, pp. 28-37.

[9] Ljiljana Stojanovic, Steffen Staab, Rudi Studer, "eLearning based on the semantic web", WebNet2001 - World Conference on the WWW and Internet.

[10] Rosic, M.; Stankov, S.; Glavinic, V., "Application of semantic web and personal agents in distance education system", Electrotechnical Conference, MELECON 2002. 11th Mediterranean,Volume , Issue , 2002 Page(s): 542 – 546, 2002.

[11] Hatem, Muna S.,Ramadan, Haider A.,Neagu, Daniel C., "e-Learning based on context oriented Semantic Web", Journal of Computer Science 1(4):500-504,2005.

[12] Sylvain Dehors, Catherine Faron-Zucker, "Reusing Learning Resources based on Semantic Web Technologies", ICALT 2006: 859-863.

[13] Bittencourt I.I., Costa, E., Soares, E. and Pedro, A., "Towards a New Generation of Web-based Educational Systems: The Convergence between Artificial and Human Agents", IEEE multidisciplinary education Magazine, vol. 3, March 2008.

[14] Gruber, T. R., "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, 5(2):199-220, 1993.

[15] Thomas Gruber, "Ontology (Computer Science) - definition in Encyclopedia of Database Systems", Ling Liu, Tamer M Özsu (Ed), Encyclopedia of Database System (2008)

[16] Erol Bozsak, Marc Ehrig, Siegfried Handschuh, Andreas Hotho, Alexander Maedche, Boris Motik, Daniel Oberle, Christoph Schmitz, Steffen Staab, Ljiljana Stojanovic, Nenad Stojanovic, Rudi Studer, Gerd Stumme, York Sure, Julien Tane, Raphael Volz, Valentin Zacharias, "KAON – Towards a Large Scale Semantic Web". Proc. of the 3rd Intl. Conf. on E-Commerce and Web Technologies (EC-Web 2002), pages 304–313, 2002.

[17] Janez Brank, Marko Grobelnik, Dunja Mladenić, "A Survey Of Ontology Evaluation Techniques", In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD) 2005

[18] Jan Brase and Wolfgang Nejdl, "Ontologies and Metadata for eLearning", pages 579–598. Springer Verlag, 2003.

[19] The Dublin Core Metadata Initiative, http://dublincore.org

[20] Learning Technology Standards committee of the IEEE: Draft Standard for Learning Object Metadata, July 2002, IEEE 1484.12.1-2002, http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

[21] Amal Zouaq and Roger Nkambou, "Building Domain Ontologies from Text for Educational Purposes", IEEE Transactions on Learning Technologies, vol. 1, no.1, January-March 2008

[22] C. Brooks, J. Greer, E. Melis, and C. Ullrich, "Combining ITS and eLearning Technologies: Opportunities and Challenges," Proc. Eighth Int'l Conf. Intelligent Tutoring Systems (ITS '06), pp. 278-287, 2006.

[23] Neches, R., Fikes, R. E., Finin, T., Gruber, T. R., Patil, R., Senator, T., & Swartout, W. R. "Enabling technology for knowledge sharing". AI Magazine, 12(3):16-36, 1991.

[24] Navigli, Roberto and Velardi, Paola, "An Analysis of Ontology-based Query Expansion Strategies", In: Proceedings of Workshop on Adaptive Text Extraction and Mining at the 14th European Conference on Machine Learning, 2003.

[25] Antonio De Nicola, Michele Missikoff, and Frederica Schiappelli, "Towards an Ontological Support for eLearning Courses", R. Meersman et al. (Eds.): OTM Workshops 2004, LNCS 3292, pp. 773–777, 2004.

[26] Alexander Maedche and Steffen Staab, "Ontology Learning for the Semantic Web", IEEE Intelligent Systems and Their Applications, Vol. 16, No. 2. (2001), pp. 72-79, 2001.

[27] James Hendler, "Agents and the Semantic Web," IEEE Intelligent Systems, vol. 16, no. 2, pp. 30-37, Mar./Apr. 2001, doi:10.1109/5254.920597, 2001.

[28] Harith Alani, "Position Paper: Ontology Construction from Online Ontologies", WWW 2006, May 23–26, 2006, Edinburgh, Scotland. ACM 1-59593-323-9/06/0005, 2006.

[29] Zhaohui Wu, Yuxin Mao, and Huajun Chen, "Subontology-Based Resource Management for Web-Based e-Learning", IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 6, June 2009.

Actual and Future Trends of the E-education and Its Costs Specification in the Czech Republic

M. Hedvicakova, P. Maresova
Department of Economics
Faculty of Informatics and Management, University of Hradec Kralove
Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic

***Abstract* – The contribution is focused on the specifics of distance education in the Czech Republic. First, reasons of emergence of distance education in the Czech Republic and abroad are analyzed. Further, two possible directions of development of e-education in the Czech Republic are outlined, where the first direction represents an expansion of capacities of the existing full-time courses. The second possible direction consists of customization of educational courses according to specific needs of students. The contribution includes definitions of individual participants of the educational process, typology of costs and revenues and break-even point modeling.**

I. Introduction

With the arrival of new technologies and varied demands in society, universities must continuously seek new ways to combine advantages of modern technologies with requirements of target customer groups. Requirements can be for example: demographic, locality, time or costs etc. Virtual universities offer the way to satisfy these requirements while continuing to multiply advantages of information and communication technologies and thus they ensure a greater return of realised investments and thereby also the growth of a university as such.

There are already many virtual universities in the world. Especially US universities have become pioneers of this type of education for the reason of their strong financial position and a high level of technological equipment. Also e-education in Australia has broken through with success. Interest in distance programs has been growing in these continents for the reason of high commuting distance to reach education.

State institutions lead by the European Union put a great accent on strengthening and changing of the approach to education. This can be also taken advantage of by regional universities, which can flexibly respond to the demand for job positions in the labour market. They can either focus on retraining courses for employment bureaus or to create "customized" courses for commercial organizations.

Supranational organizations and large companies have already been using e-learning education methods for training of their employees. With regard to the number of employees, they have become aware of advantages of on-line courses and they try to use them as their competitive advantage.

Currently we can observe that e-learning education has reached all areas, industry as well as energetic, financial sector, healthcare, telephone market, state sector and army. We cannot forget the school system, where a great majority of both universities and high schools, and to a lesser extent also basic schools, takes advantage of various education and instruction software.

II. Forms of Studies in Universities in Czech Republic

Currently three ways can be applied in the process of instruction: present lessons, i.e. traditional, face-to-face teaching and learning; distance education supported by ICT, e.g. e-learning, virtual learning environment; and blended learning, i.e. combination of present and distance elements.

In the present form of instruction in economic subjects teacher faces two main problems: students´motivation and interest, and the level of knowledge. The former one can be strongly influenced by the teacher, his/her personality, image, professional qualities, computer literacy, etc. And it results into ways how to solve the latter problem. Numerous suitable examples, applications from real life, new technologies supporting "boring" calculations, and last but not least didactic aspects – these are the means which can help. If the teacher is able to use appropriate methods and forms to explain the topic, present examples, practice them, motivate students to follow-up activities, organise team and project work – all the mentioned aspects may support the required features mentioned above. Implementation of new technologies in the process of instruction calls for both special equipment of the classroom and for the teacher who is able to work with them in an appropriate way. Modern software programmes neither work nor bring increasing interest and effectiveness by themselves. Only such a teacher, who is a professional in both the subject field and methodology, evokes the required activities from students. Having a PC, data projector etc. in the classroom, using PowerPoint presentations, links to Internet sources, it is a common standard these days. Various types of software are able to present graphs and their animations, calculate values and structure them to tables, everything with the aim to make the process of instruction easier for both students and teachers. Teacher's ability to use these tools appropriately and support all activities by personal interest and knowledge, these are the qualities that count for much.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_18, © Springer Science+Business Media B.V. 2010

It can be said the same goes for the distance education, but twice more. As there is no face-to-face contact between the teacher and students, the organisation is more demanding on both sides, teacher's and students´. Although the distance form of education is applied for tens of years, its main development came in last few years because of immense implementation of information and communication technologies into all spheres of life. The adequate use of the technologies must be emphasized, just because of the absence of direct contact. The quality of all used tools and methods, including constant motivation, strongly influences students´ results and success. Professional teams of teachers, IT workers, economists and other specialists participate in successful organisation and running the distance education.

Blended learning combines the best of the above mentioned forms, i.e. present lessons are supported by ICT, self-study activities and sources provided by the Internet are frequently used, etc. [2]

Unlike other states with a larger population or larger distance barriers (USA, Australia etc.), where the tradition of taking advantage of various education methods is greater, the market with virtual education is still only developing in the Czech Republic.

Only few Czech universities offer a compact study program by means of the e-education method. These are for example the University of Economics, Prague, offering a follow-up master's two-year program Economics and Management see [3], or VŠB – Technical University of Ostrava, Faculty of Economics, offering the study field Quantitative Management Support [4] in its follow-up master's studies. Combined programs can be found in many universities including UHK.

In our republic, the transition to this new type of education has been stimulated above all by insufficient capacities of universities for full-time studies, economic conditions of students (studying while employed), pressure of the European Union on the growth of the number of university educated inhabitants etc. Another important motive is increasing attractiveness of studies by combining study programs of individual universities, modernity of the topic of information society etc.

Also university networks have been emerging experimentally (e.g. project of Inter-university studies IUS at the universities: University of Hradec Kralove, University of Western Bohemia in Pilsen and University of Tomas Bata in Zlin, Technical University of Ostrava) [5]. Or Erasmus Virtual Economics & Management Studies Exchange (EVENE) – the international project EVENE is supported from resources of the European Union within the program eLearning and it provides students with the possibility to study at foreign universities as well. [6] Nine international universities are involved in the project.

III. Two Paths for Universities

One of the paths for universities is the above mentioned distance studies as another form of education. E-learning can be applied as an effective support of all forms of studies. The advantage for the existing "stone universities" is elimination of space and staff inadequacy of today's education. In consideration of the fact that there are no time or distance barriers, interest in this form of studies has also been growing among new groups of students (employed students, handicapped, women on maternity leave, prisoners and the like). People have also been choosing this form of studies for the reason of interactivity of courses.

The advantage for the university itself lies in a strong background, high-quality professional staff and tradition of a university as an educational institution. Another benefit for the university is reinforcement of its image as a holder of new modern technologies and a wide spectre of courses.

As regards newly emerging virtual universities, the universities which get involved in the project exchange the subjects not offered by their faculty and thereby expand the circle of their students.

The second path for universities is to create commercial e-learning courses. These are courses customized for individual companies or based on needs of the market or labour bureaus. A greater accent is put on economical aspects in these courses. Another effect is interconnection of universities with needs of the given region and therefore again reinforcement of the image of university as a core of erudition.

IV. Frameworks for Costing

If we want to calculate costs of a course, first we have to define key types of costs.

1. Categorisation of E-learning Costs

In the process of comparing costs and revenues in the distance and traditional education, a clear definition of the basic terms is necessary, i.e. what they consist of. Not the whole virtual university is evaluated, but only one distance educational course.

E-learning study costs

The standard structure for cost and revenue classification is used as it is in the common sphere of commerce, i.e. the variable and fixed costs will be counted.

a) Fixed costs

- Designing study materials – when creating an e-learning course, at the very beginning study materials must be prepared, including their electronic form in the e-environment. This is a non-recurring action.
- Authors´ royalties for creating study materials in the text form – authors must be evaluated in terms of finance.
- Printing study materials
- Costs of designing an e-course

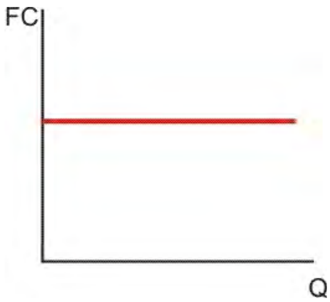


Fig. 1: Fixed costs

Stair-step function fixed costs

The level of fixed costs depends on such criteria as a number of students, frequency of tutorials, etc. It means stair-step function costs must be added to classical fixed costs, see the following examples:

- Requirements for SW and HW infrastructure of the distance education depend on the number of competitive working students.
- Financial means necessary for tutorials depend on the number of students and frequency of tutorials.

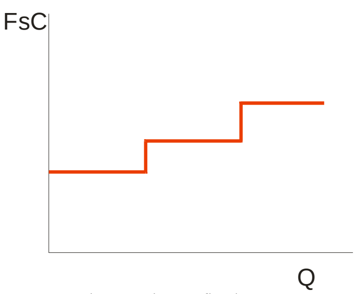


Fig. 2. Stair-step fixed costs

b) Variable costs

- Virtual environment user license
- Running an e-course – tutors and administrators of each group are evaluated
- Consumer material costs
- Rental service costs paid for tutorial rooms
- Printing study materials
- Staff costs

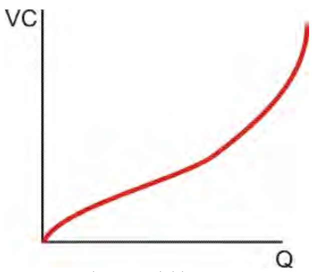


Fig. 3. Variable costs

c) Total costs

The total costs cover variable and fixed costs.

TC = FC + VC

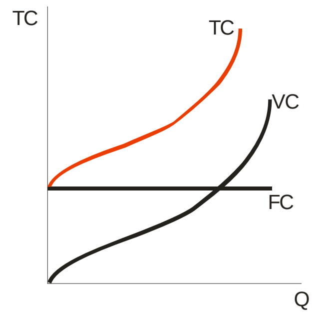


Fig. 4. Total costs

d) Average costs

Furthermore, expenses per one student are calculated, which is an aliquot part of the total costs, i.e. the total amount divided by number of students (N).

AC = (FC + VC) / N

e) Average fixed and variable costs

Average fixed and variable costs are counted in the same way:

AFC = FC / N

AVC = VC / N

2. Profits

The traditional and distance course revenue varies and consists of two items:

a)Contribution on each student (either from a state or public institution supporting education).

b) Contribution from each student (some universities require full or partial fees from students).

Total of these sums creates the income per student.

3. The Break-Even Point

Finding the rate of return comes from a principle of comparing costs and profits and finding a number of students, who have to be in the course so that the course is profitable. In economical formulation it is the break-even point. When costs and profits are equal, the more students above the minimal number we have, the higher profit we will receive.

To be able to find the point we must formulate cost and profit functions.

4. The basic cost function

The cost function reflects variable and fixed costs.

Fixed Costs:

F_1 - designing study materials

F_2 – authors royalties
F_3 – print
F_4 – creating e-course

Variable Costs:
V_1 – virtual learning environment license
V_2 - running an e-course
V_3 – consumer material
V_4 – rent for tutorial rooms and office rooms
V_5 – rental services for tutorial rooms (including HW, web camera, data projector, flip board, internet…)
V_6 – indirect expense
V_7 – staff costs
Staff costs consist of:
- Staff costs for designing modules
- Staff costs for creating supportive study materials
- Staff costs for teaching the module (tutor's salary)
- Staff costs for instruction support (study department)
- Staff costs for management
- Staff costs for examination board, etc.

N – Number of participants in a course

$$TC = (F1 + F2 + F3 + F4) + (V1 + V2 + V3 + V4 + V5 + V6 + V7) * N$$

The basic costs function (1)

Other elements could be the costs which were not displayed here, or are still unknown, and marginal to the educational institution.

The costs function taking into account step function fixed costs

Fixed costs are on the same level without consideration of the number of students. It changes by step. E.g. when the number of students increases by 25 students, another classroom for tutorials must be rented, another HW and new SW licenses bought.

Other costs are added to the cost function (1):
Fs_1 = step function fixed costs for tutorials
Fs_2 = step function fixed costs for SW
Fs_3 = step function fixed costs for HW
It results in a new function:

$$TC = (F1 + F2 + F3 + F4) + (Fs1 + Fs2 + Fs3) + (V1 + V2 + V3 + V4 + V5 + V6 + V7) * N$$

The new costs function (2)

The cost function in a time-period

The point of time has to be taken into account in the situation. The costs increase in relation to the time-period. Three time periods are used for displaying the costs.

Short course (not longer than 14 days)

In a short course no tutorials are taken into account.

The process of providing certificates is expressed as a fixed item.

F_5 – certificate costs are added to the function.

$$TC = (F1 + F2 + F3 + F4 + F5) + (Fs1 + Fs2 + Fs3) + (V1 + V2 + V3 + V4 + V5 + V6 + V7) * N$$

The costs function for short course (3)

One-term course (not longer than 6 months)

If the course lasts 1 – 6 months, tutorials must be organised. The costs for 3 tutorials are included: an introductory, middle, final one.

Zlamalova [9], Rumble [12] and other authors emphasize the necessity of introductory tutorials for courses longer than one month.

This item is a fixed cost and includes tutor's salary, fare, rent for classrooms, SW and HW, refreshments etc., and F6 – tutorial costs, are added, including F5 – Certificate.

$$TC = (F1 + F2 + F3 + F4 + F5) + (Fs1 + Fs2 + Fs3) + (V1 + V2 + V3 + V4 + V5 + V6 + V7) * N$$

The costs function for one-term course (4)

Full study (Bachelor's degree programme and Master's degree programme)

The full study is the most important and represents the most expenditures.

This type of study includes not only tutorials but also summer school (one week per 6 moths) to form those demanded skills which cannot be formed via distance instruction. The frequency of tutorials and/or summer schools is expressed by variable costs and the item "summer school" is added to the costs. Its coefficient depends on the support of special aids, laboratories, time view etc. and results in:

$$TC = (F1 + F2 + F3 + F5 + F6) + (Fs1 + Fs2 + Fs3) + v * (V1 + V2 + V3 + V4 + V5 + V6 + V7) / * N$$

The Cost function for full study (5)

Furthermore the costs covering the final phase of study, i.e. final exam and certificate ceremony, are added. The function includes F7 function – final study fixed costs.

$$TC = (F1 + F2 + F3 + F5 + F6 + F7) + (Fs1 + Fs2 + Fs3) + v * (V1 + V2 + V3 + V4 + V5 + V6 + V7) / * N$$

The Cost function for full study (6)

5. The revenue function

As presented above, it also respects the revenue particularity.

R_1 – contribution on each student
R_2 – Contribution from each student
R_3 – revenue from grants

N – Number of students

$$TR = (R1 + R2) * N + R3$$

The revenue function (7)

6. The break-even point

The break-even point introduces the place of the equilibrium between costs and revenues, i.e.:

$$TR = TC$$

or

$$(R1 + R2) * N + R3 = (F1 + F2 + F3 + F4) + (V1 + V2 + V3 + V4 + V5 + V6 + V7) * N$$

The Break-event function (8)

This is the general equation (8) of the break-even point. It arises from precondition that the whole study is a distance study and there are no personal sessions. If there are any tutorials or certificate ceremony present in a course, the appropriate costs must be included to the function.

$$\begin{aligned}&(R1 + R2) * N + R3 = \\ &(F1 + F2 + F3 + F5 + F6 + F7) \\ &+ (Fs1 + Fs2 + Fs3) \\ &+ v * (V1 + V2 + V3 + V4 + V5 + V6 + V7)/ * N\end{aligned}$$

The Break-event function with tutorial and diploma (9)

The function (9) contains tutorials during the term, summer school and certificate ceremony. Other items of the educational organisation can be included.

The break-even point depends not only on costs but also on the amount of determined fees for a course, that is for example in the case of a low fee, 120 students at least have to be in the course, but in the case of a high fee, e.g. only 25 students are necessary. Another dependency of the break-even point is that each course is subject to different costs (e.g. craft, artistic, music courses require a small number of students in the group and will be thus subject to higher costs).

Also competition in the market has been constantly growing, which is reflected in the resulting price of the course.

No less important is the image of the educational institution/university. Guarantee of a high-quality education at a prestigious educational institution will also have an impact on the price of the course.

It is further also necessary to determine a minimum number of students, on which will depend whether the course will be realised. This number will differ according to the type of the educational institution and also according to the range of its offer. [11]

V. The Process of Modelling of the Break-Even Point

The above mentioned analyses show that finding the break-even point (point of equality of costs and profits) is a question with many variables.

- Various kinds of costs,
- Various types of costs (fixed, variable),
- Various types of courses,
- Various lengths of duration of courses.

In practice, we could hardly find two identical courses at universities at this moment, which could be placed into the same group and for which the same figures would apply. Nevertheless, there is still a need (which has been increasing continuously in my opinion) to identify costs and profits of an e-learning course as precisely as possible, before the realisation of the course is initiated (in fact, it is a certain form of a business plan, whether to initiate the realisation of the course or whether to cancel it already in the phase of preparation).

Currently there is a new trend, that an increasing number of universities have been buying already prepared courses from prestigious educational institutions. As regards these courses, it is then much easier to express the amount of their costs.

I consider the above mentioned equations as a starting point for construction of an "interactive" model, that could be – after certain "user modifications" - applied in the general environment of traditional educational centres

As for me, user modifications mean the possibility to enter, before the calculation of the break-even point, a group of data into the model that describe costs of the immediate environment of the course such as e.g.:

- administrative requirements
- physical spaces
- technical background
- requirements of tutors and people directly or indirectly participating in the education (salary costs).

In my opinion, interactivity is a certain guide (some software program), that would ask in steps (or by one-time filling-in of a table) about basic variable facts related to the course that is being calculated.

This model would presuppose a certain database of realised e-learning courses (of course with their exact classification into the above mentioned groups) and on the basis of answers, it would evaluate the financial requirements and it would recommend the amount of "tuition" for the given course.

Another possible element of this questioning would be searching for the most similar courses and enabling contact with parent educational centres that realised these courses to share their experiences, so that we can avoid problems and become aware of what can be done better in the case of the given type of the course.

This all would lead to a relatively exact determination of financial requirements of the course and thus of the corresponding needs of financial resources.

The main objective of such an interactive model is to increase markedly the effectively of the financial resources expended on the distance education and to achieve such a state, when a maximum of collected resources would be invested into an effective functioning of the educational system and a minimum of resources would be expended on the bureaucratic machinery, that represents an "unnecessary evil". An increasingly higher number of educational courses have been appearing that, after the end of the subsidy program, "suddenly loose" their financial benefit for the educational institution (the course is loss-making after the end of the subsidy period) and represents a financial burden in the form of negative effectively for the institution. Correct setting of the costs and profit side of the course already in its beginnings is a necessary key factor for a long-term effective functioning of the distance education. [11]

VI. CONCLUSIONS

Having found the break-even point, the number of students is set to create and run a profitable course. The effectiveness of any course can be calculated and compared by using this simple method.

Granted finances were not taken into account in the example, but they are often the principal starting point which enables the work.

The contribution shows that any profitable and effective distance module requires great numbers of participants, according to foreign researches: at least 300 students to cover the costs (including fees and payments for printed and multimedia study materials). Correct and general information on the provided possibilities is essential for the system approach to virtual university development.

An advantage of distance courses, in contrast with the classic courses, is the fact that average variable costs, but above all average fixed costs, are decreasing in the long term. With increase in the number of participants of the course, inputs such as the number of teachers do not have to be increased in the long term: it is not necessary to rent new classrooms, it is not necessary to print other materials. All materials are created in an e-learning course that can be easily updated and students can print current information as needed.

The division of costs and their amount differ in various institutions. It is not possible to create and present a static division but a dynamic model which could be adjusted to the conditions of an institution.

Working in a virtual learning environment brings advantages to universities, e.g. decreasing costs on operation, travelling expenses, accommodation, effective training, continuous updating, etc.

Even though the distance and virtual study is running in some large and rich states (Great Britain, France, Germany, Ireland), it is financially demanding and cannot work without considerable state grants and support. The cost calculations cover author's remuneration, salaries, distribution costs, rental service costs, etc. Starting costs which may include establishing (building) administrative and study centres, preparing LMS, are very high. That is the reason why the virtual study is supported from several sources (foreign supportive programmers, tuition, production of teaching aids for other customers, selling the study materials, etc.) [7], [8], [11].

REFERENCES

[1] T. ARCHIBALD, M. NIELSEN, T. TORY NEIWERT, K. OLIVEK, Y-S. LEE, L. MARLER, *Virtual Universities: A Current Snapshot*, Utah State University URL: http://www.kwadeoliver.com/it-portfolio/research-writing/Virtual%20Universities.pdf [online], [2006-11-8].

[2] M. HEDVICAKOVA, I. SIMONOVA, *Contemporary Ways to Teaching Economic Subjects,* 38th IGIP Conference, September 6 - 9, 2009, Graz, Austria – in press

[3] The University of Economics, Prague, URL: http://www.vse.cz/kategorie.php3?IDkat=722, http://distancni.fph.vse.cz/ [online], [2008-2-4].

[4] VSB TUO, Technical University of Ostrava, URL: http://moodle.vsb.cz/moodle/206, [online], [2008-2-2].

[5] Project UIS, URL: http://www.uhk.cz/fim/studium/1188, [online], [2009-2-2].

[6] Project EVENE, URL: http://www.uhk.cz/fim/evene/2102 [online], [2008-3-11].

[7] H. Zlámalová, *Úvod do distančního vzdělávání,* CODV UP Olomouc, Andragogika, 2001, ISBN 80-244-0276-9

[8] Bakešová, M.., Koudelka, F..: Ekonomika distančního vzdělávání (Economic of Distance Education), CODV Olomouc 2001, ISBN 80-244-0274-2

[9] M. KADAVOVÁ, *Virtual Education Relation to Economic Profit of Educational Subjects*, In E-Activities: Networking the World, Proceedings of the 6th WSEAS International Conference on E-Activities (E-Learning, E-Communities, E-Commerce, E-Management, E-Marketing, E-Governance, Tele-Working / E-ACTIVITIES ´07), Puberto De La Cruz, Tenerife, Canary Islands, Spain, December 14-16, 2007, s. 384 – 388, ISBN: 978-960-6766-22-8, ISSN: 1790-5117

[10] M. KADAVOVA, A. SLABY, F. MALY, *Key Factors Involving the Design of the System of Virtual University*, In 7th WSEAS Int. Conf. on APPLIED COMPUTER & APPLIED COMPUTATIONAL SCIENCE (ACACOS '08), Hangzhou, China, April 6-8, 2008, Published by WSEAS Press, s.678 – 683, ISBN: 978-960-6766-49-7, ISSN: 1790-5117

[11] M. KADAVOVA, (M. HEDVICAKOVA) Identifying the Break-Even Point in Distance Courses, In *WSEAS TRANSACTIONS on ADVANCES in ENGINEERING EDUCATION*, Issue 5, Volume 5, May 2008, WSEAS Press, s. 282 – 294, ISSN: 1790-1979

[12] G. RUMBLE, The costs and costing of networked learning, JALN volume 5, Issue 2 – September 2001, Available at http://www.sloan-c.org/publications/jaln/v5n2/v5n2_rumble.asp [online], [2008-1-5].

Modern Improvements in the Digital Logic Laboratory

K. M. Noga, M. Radwanski
Gdynia Maritime University
Morska str. 83
81-225 Gdynia, Poland
jagat@am.gdynia.pl
marrad@gazeta.pl

***Abstract*-This paper presents examples of application for several models of physical objects controlled by programmable logic devices (PLDs) to digital logic laboratory course. During laboratory sessions students create control systems for models and realize particular assignments with application of PLDs. With the availability of new electronic components, equipment in laboratory has to change. Exercises based on standard digital integrated circuits made in TTL or CMOS technology were removed from teaching curriculum. Now the purpose of laboratory sessions is to familiarize students with programming in VHDL language and software environments designed for PLD.**

I. Introduction

Necessity of adding PLDs into digital logic education curriculum appeared together with growing number of implementations and popularity of these devices. Digital control of various devices is common today. Control using PLDs is new technical solution supplementary to used till now control systems based on microprocessors and programmable logic controllers (PLC). In comparison to microprocessor systems, programmable devices contribute simplification of some control tasks and technical realization of control system. For example, modern PLDs offer lot of universal inputs and outputs, where propagation time of signals is exactly fixed. Microcontrollers have limited amount of inputs and outputs (ports) where computation time is hard to determine, especially with utilization of high-level programming languages. Programmable logic controllers are the best solution in control of mechanical devices in heavy industry.

PLDs find more and more practical applications. Today they are generally used in data transmission as coders and decoders of transmitted signals. The new solution is implementation of PLDs in control of power electronics devices such as voltage converters, inverters and rectifiers. Control of these devices requires realization of specific functions in fixed time. Research and development concerning this implementation are conducted in many universities and research centers [1].

Altera and Xilinx corporations are the biggest manufacturers of PLDs. The most important kinds of existing PLDs are: Complex Programmable Logic Devices (CPLD) and Field Programmable Logic Arrays (FPGA). Both companies offer wide range of mentioned kinds of devices, and regularly introduce to market newer models with improved capabilities. The manufacturers of PLDs offer also free of charge basic versions of computer software intended to develop projects implemented in PLDs. Altera offers MAX+plus Baseline and Quartus II packets, Xilinx offers WebPack ISE packet. Besides, both companies offer education starter kits for academic use and release handbooks and guides on their websites [2, 3].

Popularity of PLDs forced changes in digital logic teaching curricula. Last years digital logic course on Electrical Engineering Faculty of Gdynia Maritime University (GMU) was conducted in form of lectures (30 hours), exercises (30 hours) and laboratory sessions (30 hours) during first and second year of studies. In new schedule another 15 hours of lectures and 15 hours of laboratory sessions intended only for PLDs were added. Special starter kits and models made in Department of Ship Automation and Altera DE2 Development and Education Board [4] are used during laboratory classes. Models described in this paper are connected to these kits. In [5] authors depicted first simple models used in digital logic laboratory, newer models are more complex, interactive and are closer to reality.

II. Laboratory Stations and Models

A. *Basic station with CPLD device*

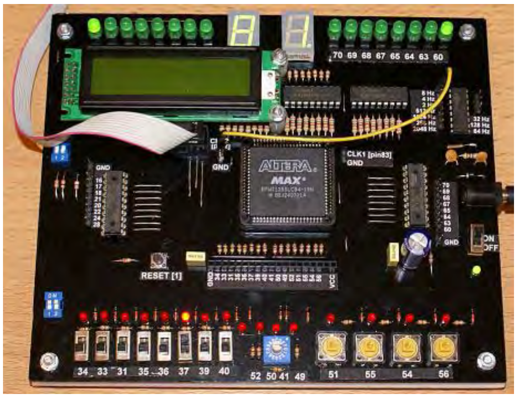


Fig. 1. Basic CPLD starter kit

The first station allowing running of programmable devices available in laboratory of digital logic at Department of Ship Automation was laboratory kit ZL1PLD from BTC. ZL1PLD kit was equipped with CPLD devices EPM7128 from Altera

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_19, © Springer Science+Business Media B.V. 2010

and XC95108 from Xilinx. There aren't universal inputs and outputs available on board, so connecting external circuits and devices to this kit is impossible. ZL1PLD kit has more technical flaws, like unavailable clock pins of CPLD. By these reasons new extended kit (Fig. 1) with Altera EPM7128 device was designed. This kit utilizes almost all capabilities of CPLD device.

B. Robotic arm

Mechanical construction of robotic arm (Fig. 2) is based on human anatomy, it contains three pivotal joints [6]. One of them is placed in the mount and allows full circle rotation, second allows vertical deviation at the base, third one allows bending at half length of the arm. At the end of the arm is placed mechanical grab, which can lift load weight up to 1 kg. Control circuit of the robotic arm consists of three units: Altera DE2 board, Atmega16-16PU microcontroller and four DC bridge controllers.

Fig 2. Complete station with robotic arm

Students during laboratory session can control this model in combinatorial mode using switches placed on DE2 board or program sequence of movements that can move various objects into selected places. The program for the FPGA device placed on DE2 board should be written in VHDL (Very-high-speed-integrated-circuit Hardware Description Language).

C. Model of dwelling house

Model of dwelling house together with devices installed inside is part of laboratory station designed for testing student projects implemented in Altera DE2 board. Model of house was built for showing possibilities of basic actuators control, it includes circuits opening automatic doors and windows and powering lights [7]. The user with DE2 board controls all devices in the house. Additionally thief alarm with movement and open doors and windows sensors is available. Air temperature sensor indicates alarm signal when temperature is higher then acceptable limit selected on setting dial.

The measurements of ground floor are 60 by 30 by 31 centimeters and 60 by 30 by 24 centimeters loft (Fig. 3). The ground floor is divided on two rooms: hall and living room. In the living room power supply, temperature setting dial and part of control electronics are placed. Near the printed circuit board is placed IDC-40PIN socket for communication with DE2 board. Entire model is shown on Fig. 4.

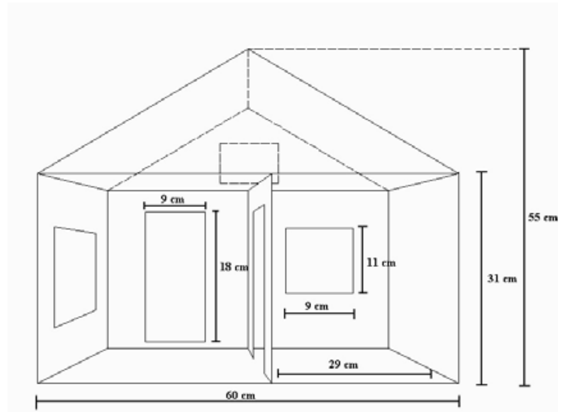


Fig 3. Model of dwelling house diagram

Movable parts of doors and windows are driven by worm-gears. Stepper motors power opening motion works. Movement stop sensors are activated when window frame is closed or fully opened.

Fig 4. Model of dwelling house – rear view

Lighting systems consists of five light sources: chandeliers in hall, living room, loft and two lamps in front of the house near the door. Chandelier in hall is turned on by manual switch or by movement sensor after detection of person in this room. The light is turned on for 5-second period and is turned off when no another movement is detected. All devices in the model can be controlled manually with use of switches and buttons on DE2 board.

Sample projects

These are some of the student projects of the control systems made during laboratory classes within digital logic course:

- assign selected switches to open windows,
- windows should close automatically after 8 seconds,

- open the door and switch on light in hall when movement detector is activated, close the door and switch of light 6 seconds after last movement,
- activate alarm when opened windows reach extreme positions.

The purpose of exercises is to acquaint students with Altera DE2 board, Quartus II environment and VHDL.

D. Model of Mobile platform

Mobile platform laboratory station (Fig. 5) consists of vehicle model (1) with control circuit, radio receiver with RS232 interface (3) and ZL11PRG (2) PLD programmer. The vehicle communicates with PC through radio receiver, RS232 interface and sends position data to application developed in Borland Delphi environment [8]. The model is based on Xilinx CPLD programmed with utilization of WebPack ISE environment.

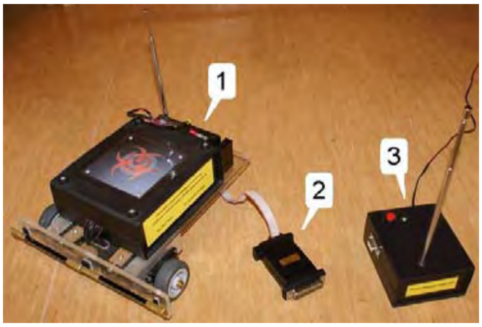


Fig. 5. Complete laboratory station

Schematic diagram of the model is shown on Fig. 6. Underbody structure is made from acrylic glass with three-wheel motion system. Control electronics and battery pack is placed on top of the platform. Three barrier sensors are placed on front bar. Rear wheel isn't driven; it works as bearing point only. Separate autonomous servomotors drive two front wheels, than platform can move in every direction. Distance is taken using wheel rotation. Rotation sensors placed on wheel axes are made from two optocouplers.

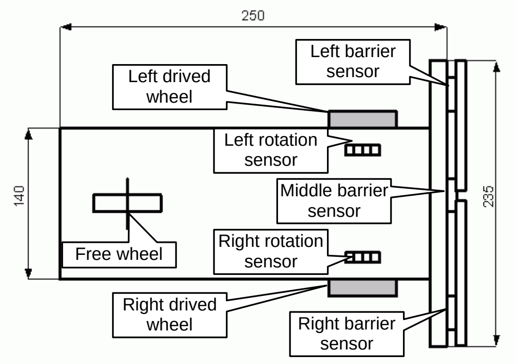


Fig. 6 Diagram of mobile platform

Eleven signals connected to CPLD are used to control this model. Four of them power the engines and set directions of wheels rotation. Next four signals are wired with rotation sensors and three last are connected to barrier sensors. Engines can be turned on and of at any moment with change of direction without pauses. Rechargeable battery Sunnyway SW613 (6V, 1.3Ah) placed behind control unit allows to operate model up to one-hour period. Data send to PC are used to draw movement trajectory with special software. Connecting characteristic direction-changing points formats trajectory plot. Position of the vehicle is transmitted to PC in fixed moments in time or when barrier is detected. Data obtained from radio receiver are decoded and stored by software in movement history file. Trajectory is stored in bitmap file at the maximum resolution of 600 by 600 pixels. Data are transmitted with 2400 bps bitrate in frames with 8 data bits and 1 stop bit.

Sample project

Student should create program that allows to move the model along fixed trajectory. Current position has to be recognized by signals in feedback from rotation sensors. Turning points should be transmitted by radio to vehicle and tracked position data should be stored in the file and visible on the screen. Student during this exercise becomes familiar with use of radio data modems for communication between PC and movable robot.

E. Multipurpose vehicle

Multipurpose remote controlled vehicle with AVR microcontroller and CPLD was built in Department of Ship Automation too [9]. Additionally vehicle can be controlled by radio with remote desktop (Fig. 7). Vehicle can run on almost any surface and may be used to penetrate hard to access or unsafe places. Load capacity about 5 kg and moderate size (36 x 23 x 8 cm) of the robot makes it perfect platform for another control and measuring devices. The main control circuit consists of EPM7128 CPLD from Altera and ATmega8-16PV microcontroller from Atmel. CPLD controls engines, proximity sensors and indicators, microcontroller controls communication through radio data modem (Fig. 8).

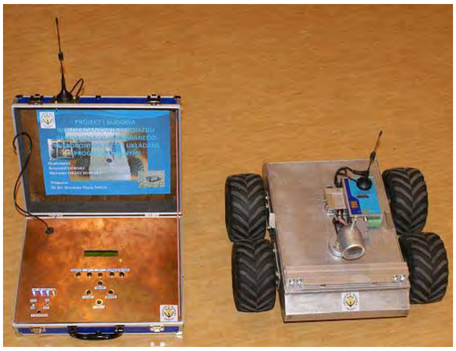


Fig. 7. Vehicle with remote desktop controller

Radio data modem with antenna, wireless camera, lights, ultrasonic sensors, temperature sensor, main power switch, battery charge socket, power and charge LED are mounted on the housing of vehicle.

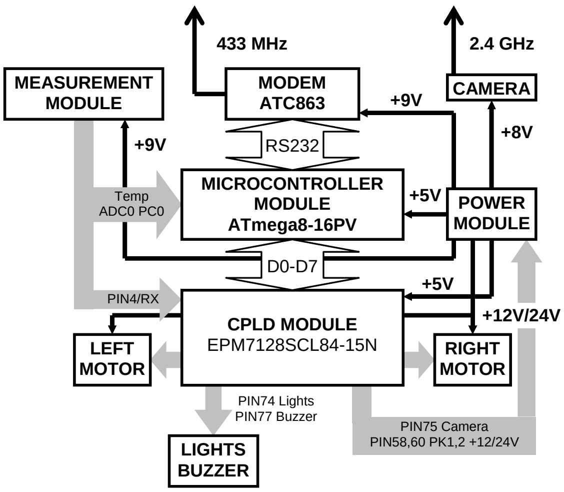


Fig. 8. Block diagram of electronic part

Vehicle motion can be controlled by portable desktop too. Desktop is based on ATmega8 microcontroller working with ATC-863 radio modem, LCD and keyboard. ATC-863 radio modem includes RS-232 and RS-485 interfaces allowing direct connection with other standard electronics devices containing UART (Universal Asynchronous Receiver and Transmitter) port.

Sample project

The vehicle should find a gap between barriers and scour room behind in autonomic mode. The task for student is to create program for CPLD, which uses proximity sensors to determine location of free space and allows sending pictures from camera. The pictures should be taken in hidden room with use of vehicle rotation to show the room in various directions.

F. Quadruped Robot

Quadruped Robot is another model designed and build in Department of Ship Automation (Fig. 9). Each leg contains two joints driven by separate servomotors [10]. Data concerning angle of flexion are transmitted in control line of the servomotor from ATmega8 microcontroller. Two kinds of steps for boost of motion capability of robot were designed. In first one robot lay between steps on the bottom side of body, in the second one robot stands on legs during entire walking cycle.

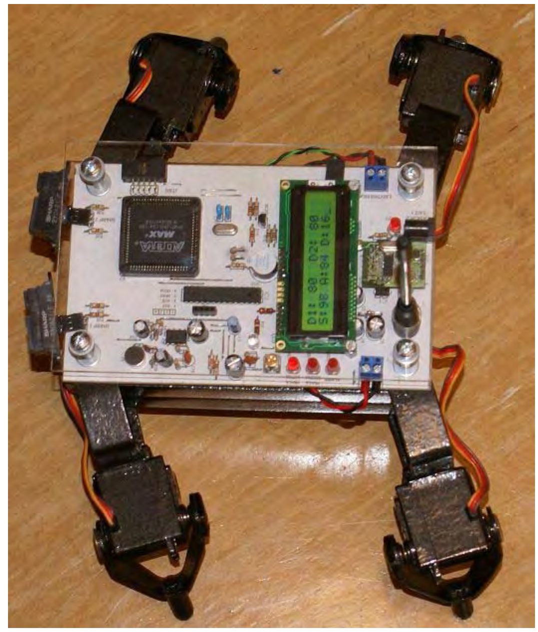


Fig. 9. Quadruped walking robot

Control circuit of the robot contains ATMega8 and ATtiny 2313 microcontrollers and EPM7128SLC84 CPLD. Robot's outfit contains sound, light, and distance sensors, battery charge indicator, buzzer and 2x16 LCD (Fig. 10).

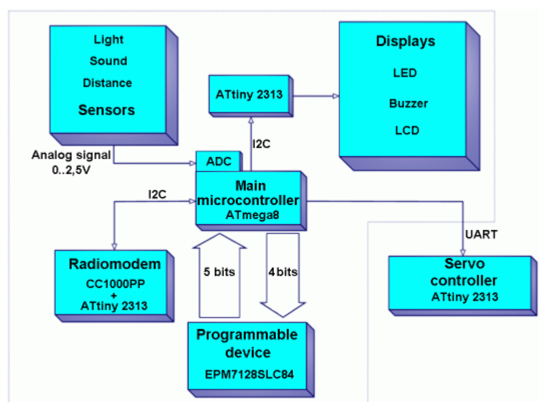


Fig. 10. Block diagram of control modules

Eight digital servomotors HX12K from HexTronik are used, two in each leg. One moves leg in horizontal plane, second in vertical. Servo controller uses simple 3-bytes communication protocol. Data are transmitted at standard UART bitrate 9600 bps. Maximum resolution of flexure depends on minimal length of pulse on output of ATtiny2313 microcontroller and equals 7.5°.

Motion control system works in three modes. In the default one running after power-on movements are autonomous according to the program implemented in programmable device. In the remote mode robot's movements are controlled by user's orders given with utilization of PC software interface. Third one is service mode allowing setting initial position of servomotors.

Robot indicates it's own status with buzzer, LEDs and LCD. Status signals are generated by ATtiny2313 microcontroller programmed with Bascom AVR environment. Radio modem CC1000PP modules are used for communication between robot and PC. Communication is based on FSK (Frequency Shift Keying) modulation, data are coded using Manchester code. Communication protocol allows bi-directional data transmission in half-duplex mode with error suppression at bitrate 9600 bps.

Sample exercise

Using VHDL and MAX+plus II or Quartus II environment student should prepare program describing sequence of movements with status indication. Buzzer, LEDs and LCD should be used to display robot state, like distance to obstacle, intensity of light or low battery level.

G. *Gantry crane*

Cargo loading is one of the most important issues in maritime transport. Laboratory station with gantry crane model (Fig. 11) allows to familiarize students not only with FPGA programming but also with logistic issues. Gantry crane model is controlled by Altera DE2 board with Cyclone II FPGA device.

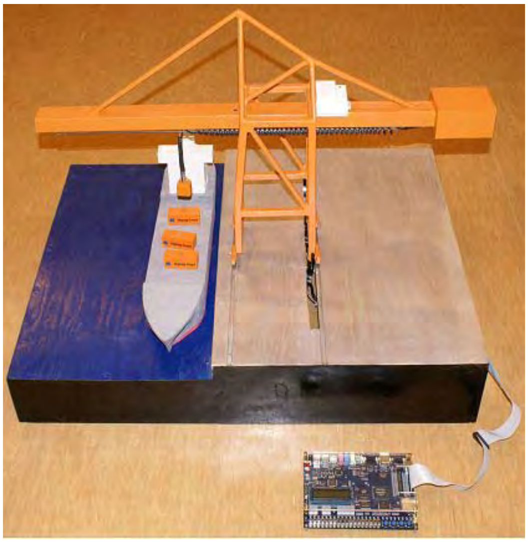


Fig. 11. Laboratory station with gantry crane model

Gantry crane model moves alongside the wharf, a trolley moves along the jib and hoisting mechanism contains electromagnet allowing lifting and carrying cargo container models between wharf and ship model [11]. Control and power system printed circuit boards and driving mechanism are located inside the mockup. Gantry crane model contains running tracks, the tower, jib, jib ties, trolley with hoisting system and hoisting block with electromagnet. Model is powered by three electric motors. DC motor moves the tower on running tracks, bipolar stepper motor placed inside the jib moves the trolley, another servomotor inside trolley lifts and lovers hoisting system.

Limitary sensors turn off motors when moving part reaches extreme position and send alert signal to FPGA device. Stepper motor is driven by controller based on ATtiny2313 microcontroller programmed in C language. Optical sensors detect positions of tower on running tracks and trolley on the jib and send sequence of pulses to FPGA device. Using this data it is possible to define current position of movable parts and to generate signals controlling next moves.

This model is based on the blueprints of Baltic Container Terminal in Gdynia [12] and is made in 1:140 scale. Basic construction of the model has measurements 165 x 145 x 530 mm, jib is 880 mm long.

Sample exercise

The most typical student task is to program sequence of actions allowing loading and unloading containers from the ship. Status of the model should be displayed on 7-segment displays and LEDs included on DE2 board.

H. *Bank vault alarm system*

To fulfill requirements of didactic occupations interesting model of bank vault alarm system was built too. Model controlled by DE2 board equipage consists of professional sensors, GSM cell phone, door opening drive and alarm system (Fig. 12).

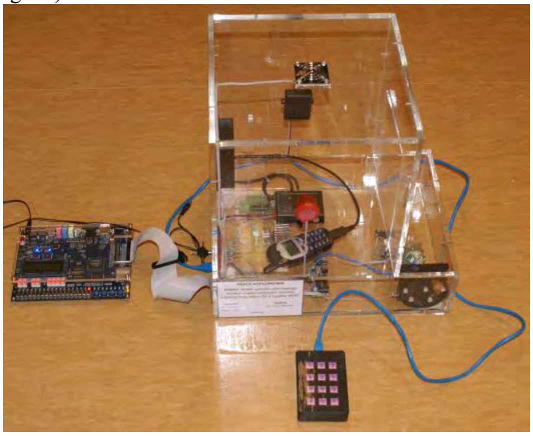


Fig. 12. Station with bank vault model

Signals from vibration sensor, microwave sensor, PIR sensor (detects changes in infrared radiation), limitary sensors, external keyboard and signals connected to actuators (stepper motor, beacon light, hooter, GSM cell phone and fan) are connected through printed circuit board with amplifiers and separators to 40-pin socked compatible with Altera DE2 board

[13]. The door is opened and closed by stepper motor with smooth regulation of speed.

Sample exercise

Using Quartus II environment and VHDL the designer should develop a project that supports vault control system, e.g. allows access after entering proper code from the keyboard. In case of double error in code system should generate alarm signals and send alert SMS through cellular network to selected phone number.

I. Control of stepper motors

Stepper motors are in common use in automatics. The method of control of stepper motors forces use of digital logic in practice. Laboratory station (Fig. 13) contains basic CPLD kit and two separate motors. Use of programmable logic allows control in all work modes of motors (full step, half step, microstepping, one-phase, two-phase and multi-phase).

Fig. 13. CPLD kit with stepper motors

Students can create projects of control systems that allow generating sequence steps in selected direction or generating fixed number of rotations. Source of the project can be entered as logical diagram or written in VHDL in MAX+plus II or Quartus II environments. This station is used during basic and advanced digital logic course.

J. Control of the plotter

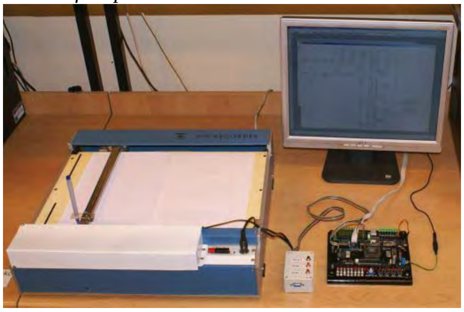


Fig. 14. Laboratory station with plotter.

Plotter is one of the oldest devices purposed to print computer graphics. Simple method of control requires only basic knowledge from the range of digital logic. Printing task is based on sending sequence of two 4-bit numbers to the plotter. These numbers determine position of the pen on X-Y plane, one additional bit represents writing or high position of the pen. EPM7128 device and basic CPLD starter kit fulfill role of the controller (Fig. 14). Students can draw simple graphics (letters or logos) containing segments of straight line, there is no possibility of drawing angles. Resolution of the drawing equals to 16 x 16 points. Control circuit can be entered as logical diagram with standard TTL 74xxx devices or as text file written in VHDL.

III. Conclusions

Models and boards with programmable logic devices presented above vary standard laboratory sessions and allow watching directly results of control process. Flexibility of PLDs gives us possibilities to build various kinds of models. Knowledge about control of DC, AC and stepper motors, using signals generated by sensors and communication between devices is required in professional career of electrical engineer.

Diversity of laboratory stations allows accurate assessment of student's work and knowledge from the wide range of engineering. Students have to be familiar not only with digital logic, but must know basics of mechanics, electrotechnics, electronics, data transmission and signal processing. Developing interactive project of control system for complex laboratory station is effortful, but students receive guides containing several examples of projects. Hard-working and ambitious persons have a chance to prove their abilities by presenting complex project with optional solutions.

References

[1] M. Baszynski, "A model of the three-phase bridge rectifier with sinusoidal source current using FPGA implementation", *Przeglad Elektrotechniczny*, vol. 3/2009, pp. 36-41.

[2] Altera Corporation website: http://www.altera.com.

[3] Xilinx Inc. website: http://www.xilinx.com.

[4] Altera Corporation, *DE2 Development and Education Board User Manual*, electronic version, 2006.

[5] K. M. Noga and M. Radwański, "Our Stations for Teaching Programmable Devices", *Innovations in E-learning, Instruction Technology, Assessment and Engineering Education*, pp. 268-273, Springer 2008.

[6] D. Kupiec amd S. Murawski, *Budowa modelu ramienia robota oraz implementacja algorytmu sterowania w srodowisku Quartus*, 1st degree thesis, GMU, Department of Ship Automation, 2008, in Polish.

[7] L. Meissinger and S. Krawczyk, *Zastosowanie ukladow programowalnych do sterowania modelem domu mieszkalnego – budowa i oprogramowanie stanowiska laboratoryjnego*, 1st degree thesis, GMU, Department of Ship Automation, 2008, in Polish.

[8] P. Ptach and, S. Sieranski, *Sterowanie wybranym modelem po okreslonej trajektorii z wykorzystaniem ukladow programowalnych*, MA thesis, GMU, Department of Ship Automation, 2008, in Polish.

[9] K. Gorski and S. Orzechowski, *Projekt i budowa wielozadaniowego pojazdu samobieznego sterowanego mikrokontrolerem i ukladami programowalnymi*, 1st degree thesis, GMU, Department of Ship Automation, 2009, in Polish.

[10] P. Wozniak and M. Wysocki, *Projekt oraz budowa kroczacego robota sterowanego ukladami programowalnymi*, MA thesis, GMU, Department of Ship Automation, 2009, in Polish.

[11] M. Witek and P. Witek, *Projekt, budowa oraz sterowanie modelem bazy kontenerowej z zastosowaniem ukladow programowalnych*, 1st degree thesis, GMU, Department of Ship Automation, 2008, in Polish.

[12] Baltic Container Terminal in Gdynia website: http://en.bct.gdynia.pl/

[13] W. Laskowski and T. Taczynski, *Model systemu alarmowego skarbca z wykorzystaniem zestawu edukacyjnego Altera DE-2 i jezyka VHDL*, MA thesis, GMU, Department of Ship Automation, 2009, in Polish.

e-Cheating and Calculator-Technology

A Preliminary Study into Casual Implications of Calculator-Technology Usage on Students' Attitude Towards e-Cheating

Zeenath Reza Khan
Instructor
Faculty of Computer Science and Engineering
University of Wollongong in Dubai
Email: zeenathkhan@uowdubai.ac.ae
PO Box 20183, Dubai UAE

Abstract

Across the globe, educational institutions are adopting e-learning tools into their curricula to cater to the ever-growing demand for technology inside the classrooms. Where traditional blackboards have been replaced by virtual whiteboards and library books by online resources, teachers are trying hard to cope with the growing competition from technology outside the classroom that has captured their students' attention and imagination, with more technology, perhaps increasing student dependency.

However, the author's previous research has shown that engaging e-learning tools to enhance student learning experience is not enough as it has negative impact on students' attitude towards e-cheating, disillusioning their awareness to cyber ethics. So, how do teachers ensure that the methods of teaching and the tools therewith truly enhance student learning and not affect their sense of ethics?

As a part of an on-going research, this paper highlights the exponential usage of calculators in the classrooms and their impact on student' attitude towards e-cheating, and recommends further studies to measure the correlation of calculator dependency in the classroom to student attitude towards e-cheating and professionalism at workplaces. .

I. Introduction

Technology in education is not a new phenomenon. It has been around for decades and has become a vital part of teaching and learning in the twenty-first century. Inside and out of classrooms, technology usage has gained popularity to feverish heights. Almost any and every technology is now used as a part of e-learning in order to enhance the overall experience for students and teachers [6]. Calculators are one such technology that users have taken for granted as a part of daily life, yet a technology that has increased student ability in all science fields, in and out of classrooms, and into offices.

This paper has been written as a subsequent resultant of an on-going research that is being carried out by the author on the casual implications of readily-available technologies on students' attitude towards e-cheating. At the start of the study, it looked as though enough research had been conducted on various aspects of e-cheating. However, as the literature review has highlighted, distinct gaps seem to exist in the literature that highlight the absence of studies showing the impact of certain technologies besides the Internet on students' attitude towards e-cheating [8]. One such technology is that of calculators that aid in computations and estimations.

This paper looks at the ease with which students can access calculators and how the usage has affected their attitude towards e-cheating.

II. Calculators now and then

The first instrument to help making calculations easy and fast was the abacus. This has wooden beds that are moved along a wire or thin wooden stick to represent numbers [9]. The young French mathematician Blaise Pascal invented the first adding machine in 1642, a device driven by gears and capable of performing mechanical addition and subtraction [1].

A revolution in calculating machines took place between the early 1960s and the late 1970s. It was during this time period that electronics for calculators was at the forefront of electronics research. 'Calculators evolved from large, expensive, mechanical machines to cheap, electronic, credit

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_20, © Springer Science+Business Media B.V. 2010

card sized devices. The development of micro-electronics for calculators was an important phase in the history of technology, which included the development of the microprocessor.' [2]

III. Calculators – Boon or Bane

Calculators are very much a part of technology that have upgraded over the last decades to become part of e-learning. As mentioned in the previous study, calculators have been a breakthrough way before computers had become common place at every home [8]. It was definitely a technology above the use of booklets with pre-calculated tables and slide rulers [10] As highlighted before, in 1965, the first pocket calculator was introduced to the market; by 1974 it had achieved providing four functions with LED screen, and although it was well over a decade before school children had their own pocket calculators, the technology advanced rapidly [10] Now-a-days, calculators range from simple scientific to graphic to programmable with large amounts of memory space, data wires to allow sharing of information, formulas and so on among friends.

However, studies have shown that where calculators make it easy for students and adults to make quick calculations, they are 'becoming a mental crutch for students, rather than a tool that promotes higher order learning' [11]. It has been seen that most often than not academic institutions ban the use of certain types of calculators in examinations to ensure students are able to work out problems upon their ability rather than with the aid of technology. Such technology in the classroom is feared to 'result in an over-reliance on technology to provide solutions, thereby stifling a student's educational and creative growth' [11]. The author has found little or no literature to show the actual effects of allowing calculators in the classroom. In many schools and universities, teachers are on high alert in examination halls, keeping an eye out for programmable calculators that students can bring in with uploaded formulas and pre-sketched graphs that would constitute cheating. But, there does not seem to be any study to actually register if there is any casual implications of allowing high-end calculators, or any computational devices that are affordable and readily available in stationary shops, on students' attitude towards cheating [11].

IV. Is it affecting students' attitude towards e-cheating?

Although the author has found no direct studies showing affects of calculators on students' attitude towards e-cheating; in the process of researching for this paper, the author has found many websites and other sources that seem to entice students into using calculators in the exam halls, for their term papers and homework in order to achieve higher grades. Sites such as Online 2 College and Teachapolis that is supposed to be a 'virtual city for teachers' are free to visit by any user sitting on a PC connected to the Internet. With such topics as 'Using Technology to Cheat • Calculator: programmable calculators can hold text, formulas, even pictures' [3] it is giving students hints on how to use the calculators in something other than simply aiding in quick calculations. Another article by Andrew Kantor in USAToday.com relents about 'add-on memory [that] lets you store software, turning your calculator into a pocket notebook… the company [Texas Instruments] is happy to point out that the TI-83 Plus is "allowed for use on the PSAT, SAT I, SAT II Math IC and IIC, AP Chemistry exam, AP Physics exam, and AP Calculus exam.' [4]. In 2005, Texas Instruments (a company that pioneered in the development of calculator technology in the 1970s) recalled thousands of calculators issued to students in the United States of America, after a student discovered that pressing certain keys converts decimals into fractions [5]. As schools expected students to have the knowledge and skill to perform this task, the calculators being used for the same purpose was labeled as cheating.

Research into existing literature has shown that calculator technology has definitely had an impact on students' attitude towards e-cheating. However, the author has found little or no evidence of actual studies carried out to establish a casual relationship between the two.

V. The Study

Based on the literature reviewed, and the corresponding study being carried out on readily-available technologies and their effects on students' attitude towards e-cheating, the following study was proposed: investigate if there are any casual implications of allowing calculator-technology in the classroom to students' attitude towards e-cheating.

To carry out this study, the sample size chosen consisted of about 100 students of mixed ethnic and educational (curricula) backgrounds. Of these, 28 were discarded as they had not completed the surveys. A survey model was developed to gather student feedback on various topics related to academic integrity and specifically to the usage of calculator and other technology. Students interviewed were chosen in random in terms of age, gender and year of program. The survey model was developed on a 5-point Likert-scale and the questions ranged from testing students' knowledge on academic integrity, to their understanding of calculators as e-learning tools and if they were using calculator-based technology in the classrooms, particularly to cheat. The results were tabulated using weighted average (ref Appendix A) for each category of question and summarized as illustrated in Appendix B. A concurrent survey was carried out on various schools to collect data on the usage of calculators in the classrooms from 2006 – 2009 alone. Table 1 in Appendix B shows the results.

VI. Results and discussion

From the study, it can be seen that when agreeing on concepts, students have a fairly good idea of what is right and what is wrong. When asked 'Academic Integrity can be defined as an adherence to a code of academic values', 67% of the students agree to this philosophy, just as 51% agree that '"E-cheating" or electronic cheating can be defined as using information technology (IT) to aid in the process of cheating in a class'. However, as has been established by previous studies [6] when it comes to applying these same concepts, students deter from their 'prior knowledge' [7]. A whopping 40.8% have agreed to the fact that they cheat on exams, 5.6% using minimized text (cheat sheets), 1.4% using mobile phones, 1.4% using memory sticks and an astonishing 13.8% using programmable calculators. Upon further interviews for clarifications, students suggested calculators as cheap and fast ways of carrying large sums of information into exam halls and using infra red to pass on the information to friends.

As illustrated by the graph below, the usage of calculators by students for various subjects have increased substantially over the three years (2006 – 2008 inclusive). At this rate of dependence, 13.8% students agreeing to use programmable calculators for cheating on exams is not a surprising percentage.

When compared to other technologies mentioned in the survey, students seem to access calculators more frequently as mode of cheating tool on exams than iPods, mobile phones, PDAs, pagers and such.

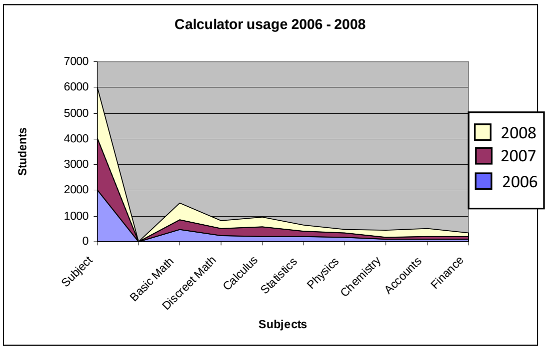


Fig 1: Calculator usage 2006 - 2008

VII. Conclusion

'Discovering, evolving, enhancing and adopting are all reasons why man moved through so many different and varied eras from ice age to the information age' [6]. However, over dependence on certain tools and technology can and have had adverse effects on the human kind before. Education and academic integrity are no different. Where teachers/parents are in a lead role to introduce and encourage student learning through whatever means available, be it traditional classrooms to electronic whiteboards, power point projectors to WebCT, these adults are also faced with the dilemma on the age-old concern of academic integrity, of testing students of what they know rather than what they can use.

Although this is a preliminary study, and a branch-out of an on-going study on the effects of readily-available technology on students' attitude towards e-cheating, the author believes it has shed some light on the possible casual effects of increased usage of calculators in the classroom by students. Where calculators have indeed saved students precious time in solving complicated calculations quickly and easily, they seem to have been added to the ever-growing list of technologies that students use in order to destroy academic integrity.

As it is beyond the scope of this paper, based on a previous study by the author, it is recommended that further investigation be carried out to measure the impact of calculators on students' academic achievements, and their ultimate usage in the work place; and study if then there is any casual implications towards e-cheating.

References

[1] Ament, P. (2007) Hand-held calculators. Troy MI. The Great Idea Finder. (Online). Available url: http://www.ideafinder.com

[2] Tout, N. (2009). Vintage Calculators. About.com. (Online). Available URL:http://inventors.about.com/gi/dynamic/offsite.htm?site=http://www.vintagecalculators.com/

[3] Bramucci, R. S. (2003). How to Cheat: techniques used by cheaters. (Online). Available URL: http://www.teachopolis.org/justice/cheating/cheating_how_to.htm

[4] Kantor, A. (2008). Cheating goes high-tech with common-place tools. USAToday.com. (Online). Available URL: http://www.usatoday.com/tech/columnist/andrewkantor/2004-05-21-kantor_x.htm

[5] Block, R. (2005). Texas Instrument calculators help students cheat. (Online). Available URL: http://www.engadget.com/2005/06/09/texas-instrument-calculators-help-students-cheat/

[6] Khan, Z. R. (2006). E-learning techniques in IT programs – are they helping my sense of cyber ethics?. Conference paper presented at the IEEE Second International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering. December 4-14.

[7] Khan, Z., Al Qaimari, G. & Samuel, S. D. 2007, 'Professionalism and ethics: is education the bridge?', in G. R. Lowry & R. L. Turner (eds), Information Systems and Technology Education: From the University to the Workplace, Information Science Reference, New York.

[8] Khan, Z. R. and Samuel, S. D. (2009). E-cheating, technologies and online sources: a critical review of existing literature. Conference Proceeding. 9th Global Conference on Business and Economics. Cambridge University, UK.

[9] Allsands.com History of Calculator. Accessed October 24, 2009. (Online). Available URL: http://www.allsands.com/History/Objects/historycalculat_sih_gn.htm

[10] Calculator.org (2009). A First Calculator. Flow Simulation Ltd. [Online] Accessed March 2009. Available URL: http://www.calculator.org/firstcalc.html

[11] Bsmarte. (2006). Managing Technology: Is technology in the classroom good or bad?. Bsmarte.com [Online] Accessed February 2009. Available URL: http://www.bsmarte.com/issues/managingtechnology.pdf

APPENDIX A

Calculating the weighted average:

In Table 1, for each Likert Item, there is a 'weight' placed depending on how close the choice is to being right.

For instance, for the question 'It is okay to download MP3 or movies from peer-to-peer websites', the weights start from 1 to 5 for each of the Likert items 'strongly agree', 'agree', neither agree nor disagree', 'disagree', 'strongly disagree' respectively. This is because if a respondent chooses 'strongly agree', he/she demonstrates he/she has chosen a wrong answer. For another question, 'Cheating can be defined as violating accepted standards or rules', the scale is reversed such that if a respondent chooses 'strongly agree', the response receives 5 point on the scale thus showing he/she has chosen the right answer. Once the weights have been placed on each response, the average is calculated by totaling response for each question and diving by the total number of respondents for each question (72 students) in each case.

APPENDIX B

TABLE 1: CALCULATOR USAGE OVER THREE YEARS FOR A VARIETY OF SUBJECTS.

Calculator Usage

Year 2006-2008

Subject	2006	2007	2008
Basic Math	468	401	650
Discreet Math	238	265	300
Calculus	210	365	365
Statistics	200	200	255
Physics	165	165	165
Chemistry	100	55	290
Accounts	110	100	300
Finance	100	100	144

TABLE 2: ACADEMIC INTEGRITY

Answer Options	**Strongly agree**	**Agree**	**Neither agree or disagree**	**Disagree**	**Strongly disagree**		
Academic Integrity is the same as Ethics	2	25	33	10	1	230.00	
Weight	5	4	3	2	1	3.19	**Weighted Average**
Academic Integrity can be defined as an adherence to a code of academic values (Webster Dictionary)	10	48	13	1	0	283.00	
Weight	5	4	3	2	1	3.93	**Weighted Average**
Cheating can be defined as an act of deception for profit to yourself	27	27	10	5	1	284.00	
Weight	5	4	3	2	1	3.94	**Weighted Average**
Cheating can be defined as violating accepted standards or rules	27	31	10	2	2	295.00	
Weight	5	4	3	2	1	4.10	**Weighted Average**
"E-cheating" or electronic cheating can be defined as using information technology (IT) to aid in the process of cheating in a class (King and Case, 2007)	14	37	16	2	2	272.00	
Weight	5	4	3	2	1	3.78	**Weighted Average**

TABLE 3: CONCEPT APPLICATION

Answer Options		Strongly agree	Agree	Neither agree or disagree	Disagree	Strongly disagree		
it is okay :								
(a) to share information among friends during tests or exams		5	10	16	29	12	249.00	
	Weight	1	2	3	4	5	3.46	**Weighted Average**
(b) to copy from the website that has the required information for an assignment		3	17	18	28	6	233.00	
	Weight	1	2	3	4	5	3.24	**Weighted Average**
(c) to copy from a text book that has the required information for an assignment		4	21	22	20	5	217.00	
	Weight	1	2	3	4	5	3.01	**Weighted Average**
(d) to write in the information from what someone else says for an assignment		1	13	19	33	6	246.00	
	Weight	1	2	3	4	5	3.42	**Weighted Average**
(e) to copy from another friend who has the information for an assignment		2	7	15	37	10	259.00	
	Weight	1	2	3	4	5	3.60	**Weighted Average**
All of the above points a-e but with due citations and reference list		28	30	7	4	0	289.00	
	Weight	5	4	3	2	1	4.01	**Weighted Average**
It is okay to install a copy write software given to me by a friend		9	25	24	10	3	186.00	
	Weight	1	2	3	4	5	2.58	**Weighted Average**
It is okay to download MP3 or movies from peer-to-peer websites		18	35	9	7	3	158.00	
	Weight	1	2	3	4	5	2.19	**Weighted Average**
It is cool to buy pirated movies from vendors on the streets for AED 5/- instead of the original for more than AED40/-		13	21	15	17	5	193.00	
	Weight	1	2	3	4	5	2.68	**Weighted Average**
'Anything goes' is a sure attitude to success		4	9	30	22	4	220.00	
	Weight	1	2	3	4	5	3.06	**Weighted Average**

TABLE 4: WHAT I PRACTICE

Cheated on exams: 40.8% (29 out of 72) have answered YES

Cheated on exams using technology such as :	Yes	No	%
A) Mobile phone wireless connectivity	1	71	1.4
B) PDAs	0	72	0
C) Programmable calculators	10	68	13.8
D) Pagers	1	71	1.4
E) Minimized text using photocopy machine to make cheat sheets	4	68	5.5
F) iPOD	0	72	0
G) e-dictionaries	0	72	0
H) Memory sticks for online exams	1	71	1.4

Perspectives on Critical Thinking through Online Discussion Forums in Engineering Mathematics

[1]S. M. Jacob
Swinburne University of Technology (Sarawak Campus), Malaysia.
[2]H. K. Sam
University Malaysia Sarawak, Malaysia.
[1]sjacob@swinburne.edu.my and [2]hksam@fcs.unimas.my

***Abstract*-This study set out to examine the online mathematical communication and measure the progress of critical thinking (CT) skills of students that occurred in collaborative groups, during two collaborative problem solving sessions via online Discussion Forums (DFs), in a first year Engineering Mathematics unit which spans 14 weeks. It was found that the calculated CT scores have improved from the first to the second DF session. The CT scores were connected with moderation techniques associated with Socratic Questioning and reflection scaffolds, during the first session, but not in the second. Students' perceptions on the DFs were analyzed and was found that majority of the students found the DF sessions quite challenging to their thinking skills and profitable for developing their problem solving skills.**

I. Introduction

Critical thinking is both a process and an outcome [1]. As an outcome, it is best understood from an individual perspective—that is, the acquisition of deep and meaningful understanding as well as content-specific critical inquiry abilities, skills, and dispositions. Judging the quality of critical thinking as an outcome within a specific educational context is the responsibility of a teacher as the pedagogical and content expert. As a product, critical thinking is, perhaps, best judged through individual educational assignments. Critical thinking (CT) and mathematical problem solving have existed side by side and have attracted research over the decades.

The research questions answered through this study, done on online Discussion Forums (DFs) for a batch of Engineering Mathematics students were: (1) What was the level of CT skills of the participants, individually and group wise, in the online DFs? (2) What was the effect the moderation techniques have on the level of CT skills in the online discussion forums? (3) What were the participants' perceptions about the online DFs and the impact of the DFs on their critical thinking skills?

II. Review Of Literature

A. Critical Thinking and problem solving

Erwin [2] defines problem solving as a step-by-step process of defining the problem, searching for information, and testing hypotheses with the understanding that there are a limited number of solutions. Critical thinking is a broader term describing reasoning in an open-ended manner, with an unlimited number of solutions. The critical thinking process involves constructing the situation and supporting the reasoning behind a solution. It focuses on the process of learning rather than just attaining information. It involves discovering how to analyze, synthesize, make judgments, and create and apply new knowledge to real-world situations. A simple standard definition of critical thinking adopted in this study is "the ability to use acquired knowledge in flexible and meaningful ways, through understanding the problem or issue, evaluating evidence, considering multiple perspectives, and taking a position" [3], p. 275.

B. Socratic Questioning

Yang, Newby & Bill [4] has used Socratic Questioning to promote critical thinking skills in asynchronous discussion forums. A powerful method of cultivating critical thinking in a group, Socratic Questioning uses the technical strategy of regressive abstraction and develops a rigorous inquiry into the ideas, concepts and values held by people with regard to the topic. The questions fall under the six categories of Socratic questioning prompts [5], namely, questions of clarification, questions that probe assumptions, questions that probe reasons and evidences, questions about viewpoints or perspectives, questions about the question and questions that probe implications and consequences. Instead of directly providing statements as clues for students to advance in their process of solving the problem, questions are posed with a reasoning approach, from which the students are led to thinking into the different aspects/stages of the problem.

C. Asynchronous Discussion Forums and CT

There are two approaches [6] to teach CT using content disciplines: (1) The "infusion" approach where critical thinking skills are taught manifestly using the discipline's content; and (2) The "embedded" approach, where critical thinking skills are taught in indirect ways without spelling it out to students. The asynchronous discussion forums are seen as an attempt to try the embedded approach. Discussion forums were identified by [7] as one of the effective forms of assessments of critical thinking. Perceptions of asynchronous communication tools have been reported by [8] and [9]. In their study on cross-age peer tutors' contributions in asynchronous discussion groups, [10] have commented that students' learning experiences and perceptions about the scaffolding offered on asynchronous

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_21, © Springer Science+Business Media B.V. 2010

discussion forums might nourish a prospective qualitative design.

III. Research Study details

The study involved the first semester Engineering Mathematics unit of the Bachelor of Engineering programme offered by the Swinburne University of Technology (Sarawak Campus) during the year 2009. Out of the total number of 115 students, 60 participants were chosen by voluntary consent. The number of participants was fixed at 60, since the study was an in-depth study involving a lot of variables, especially the analysis of the online forum postings of individual students. A similar study on online forums in mathematics done by Kosiak [11] has used around 50 participants. The participants filled in the Consent Form agreed by the Swinburne University Ethics Committee, as an agreement to their participation in the study. The participants consisted of a mixture of Malaysian (83%) and international (17%) students. The 10 international students, out of the total of 60 participants, comprised students from Indonesia (1), Iran (1), Iraq (1), Kenya (2), Mainland China (2), Saudi Arabia (2), and Sri Lanka (1). Generally the students of this unit are predominantly male. The sample reflects this typical scenario – it was an uneven mixture of males (85%) and females (15%). The first year university students go through a transition stage from school into university. Hence they need a lot of support in terms of coping up with new learning environments and taking ownership of their learning. The online mathematical forums were used as a means of establishing good rapport between the students and the lecturer, and between the students themselves. All these students were new to the experience of using discussion forums in the educational setting.

The unit was supported by the online Learning Management System (LMS) called the Blackboard Learning System (http://www.blackboard.com), or BBLS. The Discussion Forums were available on the BBLS. Two problem solving sessions were planned on these Discussion Forums - that is, ill structured problems were posted at two different times (in Week 4, DF1 and in Week 10, DF2) during the 14 week long course. The goal of these sessions was for students to work collaboratively to solve and understand the ideas on these ill-defined problems [12]. The 60 students were divided into 11 groups of 5 or 6. The first forum carried a weightage of 7 %, and the second carried a weightage of 8%. 15% of the course work marks was assigned for the Discussion forums to encourage students into active participation. The instructor encouraged these collaborative sessions by moderating (scaffolding) the discussion thread in order to stimulate both mathematical learning and mathematical understanding. In this study, during the first three weeks of the unit, when the discussion forum sessions have not yet begun, the students were given access to examples of ill structured problems and solutions. The ill structured problems were chosen by the lecturer, who is also the moderator of the forum, from a standard book of engineering application problems called "Modern Engineering Mathematics" [13]. During the course of the discussion, the instructor moderated, or, scaffolded and modeled the process through Socratic questioning.

The proposed model shown in Fig. 1 was based on the following principles: (1) Learning is a dialogical process in which communities of practitioners socially negotiate the meaning of phenomena [14],[15],16]; (2) views of knowledge are constructed by, rather than transmitted to, the learner [17]; (3) ill-structured problems and Socratic questioning are stimuli [4],[5] for critical thinking; and (4) asynchronous discussion forum is a collaborative mind tool [17] which contributes to the development of critical thinking and effective communication. The model integrates the principles of the model presented by [15] for the analysis of collaborative knowledge building in asynchronous discussions.

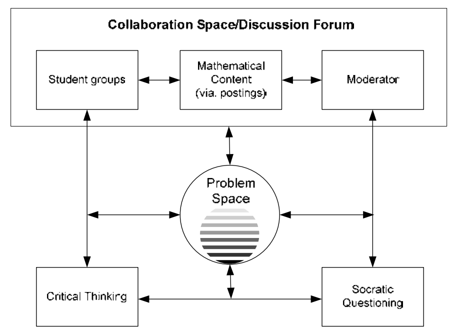


Figure 1. A proposed model for designing an online mathematical discussion forum environment.

The model conceived an ill structured problem or, its problem space as the focus of the discussion. As seen in Fig. 1, the fundamental unit of the whole activity is the problem space [18], [19] constructed from the ill structured problem by the students. The creation of a problem space involves mapping the problem statement from the ill structured problem onto prior knowledge and constructing a personal interpretation of the problem. The process of solving the problem would follow the iterative and cyclical process [19] that ill-structured problems follow. The process would expect the participants to articulate goals, verify the problem, relate problem goals to problem domain, clarify alternative perspectives, generate problem solutions, gather evidence to support/reject positions, construct arguments, implement and monitor solution. In the process, the participants generated clarification questions, assessment questions, inferences, strategies (viz., the four phases of critical thinking stated in the model shown in table 2) and took part in active and collaborative learning. The moderator would pose scaffolding questions [12] to clarify key mathematical concepts and to advance/direct the discussions towards the solution.

Data was collected in this study using two modes: (1) The Discussion Forum postings; (2) Survey Questionnaires- perceptions on discussion forums and critical thinking questionnaire.

TABLE I
SCORING TABLE FOR CT CATEGORIES

Category	No. of postings	Weightage
Clarification	1-2	1
	3-4	1.5
	>4	2
Assessment	1-2	3
	3-4	3.5
	>4	4
Inference	1-2	5
	3-4	5.5
	>4	6
Strategies	1-2	7
	3-4	7.5
	>4	8

An ill structured problem was posted on the discussion forum by the instructor, students were briefed in detail about the sessions and given the deadline of one week to solve. The postings were downloaded by the instructor after the sessions were over. The postings were coded into the four categories, using the indicators in the model as guides. In cases where more than one critical thinking process appeared within a posting, only one code was associated, which seemed to be the most important in the context.

The model for content analysis of the postings in the current study was adapted from two models- the model proposed and tested by Perkins and Murphy [20], and the framework for assessing critical thinking developed by Paul and Elder [21]. The new model is shown in Table II [22], with the indicators and description to each phase/category. The higher levels of critical thinking are associated with the stages of Inference and Strategies.

Also a weighted critical thinking score (CT1 and CT2) was associated with every student for DF1 and DF2 respectively, based on the classification of the postings, as shown in Table I. The score has been developed for the purpose of quantitative analysis based on the CT model. The table was developed with the following principles: (1) The limit of the number of postings has been set, assuming the average number of postings of a student, per category, for one discussion forum was around three. (2) Higher weights are associated with Inference and Strategies, to indicate the higher levels of critical thinking. (3) The maximum weightage of one category differs by one from the minimum weightage of the next category. Hence the maximum score possible is 20, if a student has more than four postings in all categories.

The perceptions of students were measured using the Critical Thinking and Discussion Forum Questionnaire (CTDFQ). The questionnaire had three sections – section 1 on CT, consisting of 4 questions, section 2 on DFs, consisting of 9 questions, and section 3 consisting of one open ended question asking for qualitative comments.

TABLE II
MODEL FOR IDENTIFYING CRITICAL THINKING IN MATHEMATICAL PROBLEM SOLVING SESSIONS OVER DISCUSSION FORUMS

Clarification			
Formulates the problem precisely and clearly.			
Analyses, negotiates or discusses the scope of the problem	Identifies one or more underlying assumptions in the parts of the problem	Identifies relationships among the different parts of the problem	Defines or criticizes the definition of relevant terms
Assessment			
Raises vital questions and problems within the problem.			
Gathers and assesses relevant information.	Provides or asks for reasons that proffered evidence is valid or relevant.	Make value judgment on the assessment criteria or argument or situation.	
Inference			
Reasons out based on relevant criteria and standards			
Makes appropriate deductions from discussed results.	Arrives at well thought out conclusions	Makes generalizations from relevant results.	Frames relationships among the different parts of the problem.
Strategies			
Thinks and suggests open mindedly within alternative systems of thought.			
Propose specific steps to lead to the solution.	Discuss possible steps.	Evaluate possible steps.	Predicts outcomes of proposed steps.

IV. RESEARCH STUDY RESULTS

The first Discussion Forum, DF1 occurred during Week 4, and the second Discussion Forum, DF2 occurred during Week 10 of the 14-week mathematics unit. During the course of the one week participation, the total number of posted messages in the discussion forum sessions- namely DF1 and DF2 was 616 and 539 respectively, as it occurred in the 11 groups. Fig. 2 shows the mean number of postings across the 11 groups. As seen from the Fig. 2, participation varied among groups and individuals. The descriptive statistics in table III are compiled from DF1 and DF2, hence each variable carries a '1' or '2' in its name, to denote it comes from DF1 or DF2. The arithmetic means in table 3 show that the number of postings was more in the lower phases of critical thinking, namely clarification and assessment, than in the higher phases. But the standard deviation (S.D.) and the quartiles (Q1, Q2 and Q3) reveal substantial variability in the levels of critical thinking of the students. The quartiles give a better picture of the distribution of the variables, since the distribution was not normal.

Research Question 1a: What is the level of CT of individuals in the online discussion forums? Shapiro-Wilk's test for normality done on the variables – the number of postings in the four CT categories, and the CT model scores (CT1 and CT2), showed that the data values are not normally distributed ($p < 0.05$) for all these variables. But the parametric calculations and tests were valid on the data set, since the sample size of 60 is large ($n > 30$) enough for the normal assumption of distribution of the data.

The distribution of the data for the CT categories and the CT model scores looked very widely distributed among the batch of students. One fourth of the students posted no postings in the higher phases of CT – namely, inference and strategies, but another one fourth posted more than 4 postings.

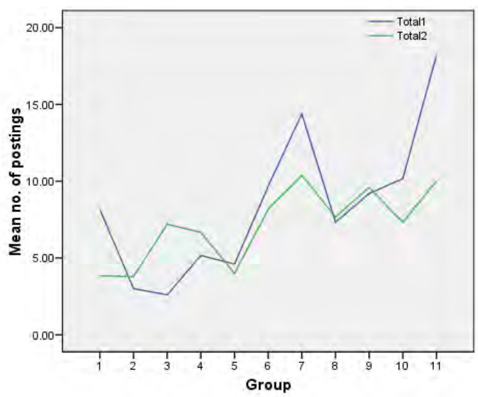


Figure 2. Line chart showing the mean total number of postings across the groups in DF1 and DF2.

TABLE III
DESCRIPTIVE STATISTICS OF THE 60 STUDENTS

Category	Mean	S.D.	Q1	Q2	Q3
Clarification1	2.77	2.580	1	2	4
Assessment1	2.93	2.622	1	2	4.75
Inference1	0.85	1.400	0	0	1
Strategies1	0.50	0.948	0	0	1
Unclassified1	1.33	1.980	0	0	2
Clasiification2	2.62	1.958	1	2.5	4
Assessment2	2.50	1.827	1	3	4
Inference2	0.77	0.909	0	0.5	1
Strategies2	0.43	0.810	0	0	1
Unclassified2	0.81	1.181	0	0	1
CT1	8.33	6.045	4	7	12
CT2	8.76	5.112	4.5	9.5	14.25

The median (Q2) shows that half of the students posted fewer (more) than 2 postings in the lower phases of CT. But half of the students have just one or more postings in the higher phases of CT. The CT scores increase from DF1 to DF2, as seen from the median scores of 7 and 9.5 respectively. The results showed that CT skills measured in the mathematics problem solving sessions were very varied. But, generally CT skills were exhibited in the lower phases of critical thinking, throughout the two discussion forums. A Pearson correlation proved significant between CT1 and CT2 ((Pearson correlation = 0.644, $p < 0.01$) at 1% significance level. Or, the CT model scores for the students move linearly between DF1 and DF2. As students progressed from DF1 to DF2, the students' CT skills exhibited in the problem solving sessions over the DF also showed an improvement.

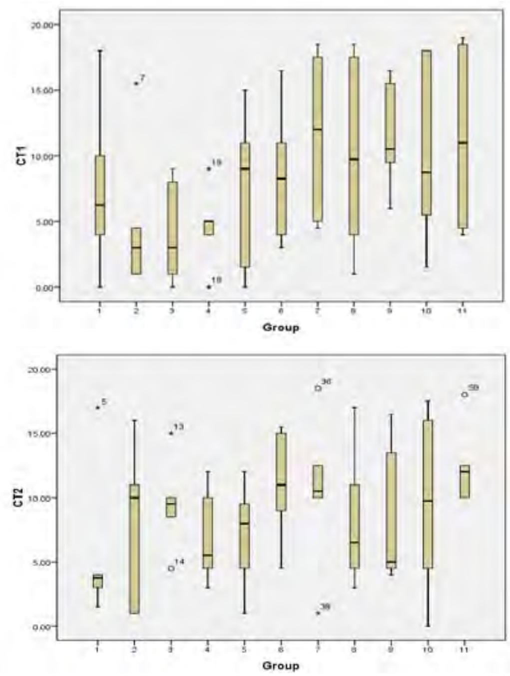


Figure 3. Box plots showing the CT1 and CT2 scores across the 11groups.

Research Question 1b: What is the level of CT of groups in the online discussion forums? The variations were noticeable within the groups, much more than between the groups. In DF1, the CT model scores looked roughly similar for all the 11 groups – lying between 2 and 18. But there were exceptional students in Groups 2 and 4. The CT model scores in DF2 show much variation within the group members, than between the groups. There was at least one student who scored well above the rest of the group members in Groups 1, 3, 7, and 11. Also there was a low scorer who was way below the rest of the group members in Groups 3 and 7. A MANOVA was employed to test for differences in the four phases of CT and the CT model score between the different groups in DF1 and DF2. Taking these phases as the combined dependent variables, the test was found significant between the groups in DF1 ($F = 1.498$, $p < 0.05$) and DF2 ($F = 1.460$, $p<0.05$). But univariate tests revealed that the significant differences were found for the phases, namely clarification ($F = 2.743$, $p < 0.01$) and assessment ($F = 2.593$, $p < 0.01$) in DF1 and only for assessment ($F = 2.173$, $p < 0.05$) in DF2. A one-way ANOVA done for the CT model scores, CT1 ($F = 1.647$) and CT2 ($F = 0.975$) between the different groups showed the differences were not significant at 5% significance level.

Research Question 2: What is the effect of moderation techniques on the level of CT of individuals in the online discussion forums? MANOVA was employed to test the effects of moderation (scaffold) on the participant's phases of CT and the CT model scores, in DF1 and DF2. The moderation types were: Type 1- Question scaffolds (Socratic questioning) only and Type 2- Question scaffolds (Socratic questioning) plus reflection scaffolds.6 out of the 11 collaborative groups, had the type 1 scaffold, and the rest of the groups had the type 2 scaffold. Thus the moderation type, which is a combination of Socratic questioning and reflection scaffolds, produced significant differences in student postings in the first forum DF1. The differences were noticed in the phases of clarification, assessment and strategies. But in DF2, students seemed to have become more independent that the postings didn't vary much between the two groups. Or, moderation seemed to have less effect as time progressed and students were used to the pattern of discussions over the DFs.

Research Question 3: What are the student perceptions on CT and DFs as they experienced during the 14 week course? Section 1, covering CT, consists of 3 questions rated on a 4-point Likert scale: Strongly Agree (1) to Strongly Disagree (4). The 4-point scale was chosen in favour of the usual 5-point scale since the researcher does not believe in "neutral" people [23] walking about with no positive or negative response towards the issue in hand. There was a fourth question asking for the student's self evaluation of his/her own CT rated on a scale of Excellent (4) to Poor (1). Student perceptions have favoured the fact that CT was understood as a result of doing the particular mathematics unit. Group work was also seen as helping CT skills. The argument that CT skills became better from DF1 to DF2 showed a mean closer to 3. Majority of the students were echoing the fact that their CT skills (as they understand) became better from Forum 1 to Forum 2. It is noticed from the results of research question 1 that the CT model scores also reflected the same result- there was seen a general progress in CT skills from DF1 to DF2.The students' rating of their own CT was an average (the mean value is close to 2) on a scale of 1-4. Majority of the students saw themselves as belonging to the class of average critical thinkers.

Section 2 of the CTDFQ dealt with the DFs and the problem solving sessions. There were 9 questions in this section – 5 positively worded statements, 3 negatively worded statements and 1 question to rate the DF problem solving sessions activated on the Blackboard Learning System. Majority of the responses were positively favouring the DF sessions as an excellent way to promote CT. They did agree to the fact that the discussion and pace of the DF sessions became better from DF1 to DF2. The students have positively agreed to the fact that their attitude to mathematics has been improved as a result of the DFs. The students have disagreed to the comment that moderation did not play an important role.

TABLE IV
SAMPLE COMMENTS OF STUDENTS

Positive comments	Negative comments
Questions connected with daily life applications boosted my appreciation towards mathematics.	Unsystematic posting of threads causes discomfort.
I get to see different opinions over one problem. Thus I find out more than one way to approach the given problem.	All group members do not think and post at the same pace; So some group members might be discouraged to post any further.
It is an easier way to get marks for the course work.	Not all members are active. Besides we sometimes could not care less about the Disc. Forums.
It's good to have it online; Saves time/hassle free.	No active participation from the group members.
We get idea from others, even if I don't know how to solve the problem.	But two Discussion Forum sessions is too much.
You explore and think outside the box. It gives a wider picture than what you see and hear in the classroom.	Lack of participation from group members is like a cold blanket, and disables us from exchanging views.
The questions are great and help engineering students to see how simple mathematics could be used in applications.	I am not a fan of team work for problem solving. I prefer to work individually than in a team.
It makes me excited, and I will not give up without getting the answers; I like critical thinking.	My group members did not co-operate well. It looks like I am discussing with the lecturer.
It increases my group work skills, critical thinking skills. Can share knowledge with friends.	The time frame is too short. I need time for solving any question because the main problem with me is how to begin.

But they have agreed that the moderator modeled by posing appropriate discussions. Nearly 40% of the students sounded that they participated simply because of the course work marks allotted for participation in the DF sessions, and that participation was a demanding task. This was reflected in some of the negative comments to the open questions listed later. But the majority of students have participated with a more positive mind, and enjoyed the online sessions which made them think out of the box. The rating of the DF sessions on the Black Board was also a pleasant mixture of 'Good' and 'Average'. But the mean score pointed to the fact that the sessions were rated good.

Section 3 consisted of an open ended question asking for positive and negative comments on the experience of students with the DFs.

Table IV shows a set of positive and negative comments made by a few students. It was clear from the responses that these students have never been exposed to such problem solving sessions in mathematics. Team work skills, innovation, sharing with friends, and the fun of being free to think beyond limitations were seen as the positive sides of the DF sessions. But, being first semester students who were new to the university environment, some of them felt that the sessions required a lot from them on top of their packed schedules. Also lack of response from their team mates, preference for individual work, short time frame were the reasons for the negativity of the students.

V. CONCLUSIONS

The idea of measuring critical thinking (CT) skills of students through the online Discussion Forums, proved to be very useful to the instructor, especially in the context of a large class (more than 50 students), where it is difficult to assess the thinking skills, personally or one to one. The instructor could classify the students' postings based on their thinking skills and thus give feedback to the students. In general, the majority of the postings fell into the lower categories of CT, as was the case noticed by Lee [1] and Kosiak [11]. But the CT skills have improved from the first session to the second, as shown by the CT model scores. But there were students who were ahead of the others, considered as outliers. The moderation process was very much necessary in the beginning stages, but during the second session, students seemed to be confident to move on by asking, assessing and discussing among the group members. The students perceived the forums as a boon to work on their CT, problem solving skills and majority of the students appreciated the innovation and challenge offered by the DF sessions. Students also gave the researchers enough feedback to work on.

REFERENCES

[1] K.S. Lee, "Effects of Individual versus Online Collaborative Case Study Learning Strategies on Critical Thinking of Undergraduate Students. Doctoral Dissertation," The University of Texas at Austin, 2004.

[2] T. D. Erwin, "Definitions and assessment methods for critical thinking,problem solving, and writing," in T. Bers, Assessing critical thinking in community colleges, *New Directions for Community Colleges*, vol. 130, pp. 15-25, 2005.

[3] S. VanderStoep and P. Pintrich, *Learning to learn: The skill and will of college success.* Upper Saddle River, NJ: Prentice Hall, 2003.

[4] Y-T.C.Yang, T.J. Newby, and R.L. Bill, "Using Socratic Questioning to promote critical thinking skills in asynchronous discussion forums in distance learning environments," *American Journal of Distance Education*, vol. 19(3), pp. 163-181, 2005.

[5] Y-T.C Yang and H-A. Chou, "Beyond critical thinking skills: Investigating the relationship between critical thinking skills and dispositions through different online instructional strategies," *British Journal of Educational Technology*, vol. 39 (4), pp. 666-684, 2008.

[6] H. Innabi, H. and O.E. Sheikh, "The Change In Mathematics Teachers' Perceptions Of Critical Thinking After 15 Years Of Educational Reform In Jordan," *Educational Studies in Mathematics*, vol. 64, pp. 45 – 68, 2006.

[7] G. E. Perry, "Critical Thinking and Web-Based Education in Maryland Community Colleges: How the Medium Promotes Development". Unpublished doctoral dissertation, University of West Virginia, in Bers, T. H., Assessing Critical Thinking in Community Colleges, *New Directions For Community Colleges*, vol. 130, Wiley periodicals, Inc., pp. 15-25, 2005.

[8] M. Guzdial, P. Ludovice, M. Realff, T. Morley, K. Caroll, and A. Ladak, The challenge of collaborative learning in engineering and math., Frontiers in Education Conference, 2001, 31st Annual, pp. T3B-24-29, 2001. Retrieved January 2008 from http://coweb.cc.gatech.edu/cs1.

[9] T. Karsenti, "Teacher Education and Technology: Strengths and Weaknesses of Two Communication Tools," Proceedings of the 2007 Computer Science and IT Education Conference, 2007. Retrieved January 2008 from http://csited.org/2007/83KarsCSITEd.pdf.

[10] M.D. Smet, H.V. Keer, and M. Valcke, "Cross-age peer tutors in asynchronous discussion groups: A study of the evolution in tutor support," *Instructional Science*, 37(1), pp. 87-105, 2009.

[11] J. J. Kosiak, "Using Asynchronous Discussions To Facilitate Collaborative Problem Solving In College Algebra", Doctoral Dissertation, Montana State University, 2004.

[12] M-S. Chiu, "Approaches to the teaching of creative and non-creative mathematical problems," *International Journal of Science and mathematics education*, vol. 7(1), pp. 55-79, 2009.

[13] G. James, *Advanced Engineering Mathematics (3rd Edition).* Pearson Prentice Hall, 2004.

[14] M. Mabrito, "A Study of Synchronous Versus Asynchronous Collaboration in an Online Business Writing Class," *The American Journal of Distance Education*, vol. 20(2), pp. 93-107, 2006.

[15] S. Schrire, "Knowledge building in asynchronous discussion groups: going beyond quantitative analysis", *Computers & Education*, vol. 46(1), pp. 49 – 70, 2006.

[16] L. S. Vygotsky, *Mind in society: The development of higher physiological processes*. Cambridge, MA: Harvard University Press.

[17] S. Schrire, "Interaction and cognition in asynchronous computer conferencing," *Instructional Science*, vol. 32, pp. 475-502, 2004.

[18] A. E. Campbell, V. M. Adams, and G. E. Davis, "Cognitive Demands and Second-Language Learners: A Framework for Analyzing Mathematics Instructional Contexts," *Mathematical Thinking & Learning*, vol. 9(1), pp. 3-30, 2007.

[19] S. Sendag, and H.F.Odabasi, "Effects of an online problem based learning course on content knowledge acquisition and critical thinking skills," *Computers & Education*, vol. 53, pp. 132-141.

[20] C. Perkins and E. Murphy, "Identifying and measuring individual engagement in critical thinking in online discussions: An exploratory case study," *Educational Technology and Society*, vol. 9(1), pp. 298-307, 2006.

[21] R. Paul and L. Elder, *Critical Thinking: Learn the Tools the Best Thinkers Use*. Prentice Hall, 2006.

[22] S. M. Jacob and H.K. Sam. "Effect of Online Discussion Forums to Improve Mathematics Problem Solving and Critical Thinking – A Comparative Study," Proceedings of the 5th International Conference on Innovations in Information Technology (Innovations 2008), United Arab Emirates, ISBN 978-1-4244-3397-1, pp. 366-370, 2008.

[23] M.G. Lodico, D.T. Spaulding, and K. H. Voegtle. *Methods in Educational Research: From Theory to Practice* [E-Book], 2006.

Digital System Description Knowledge Assessment

K. Jelemenská, P. Čičák, M. Jurikovič and P. Pištek
Institute of Computer Systems and Networks, Faculty of Informatics and Information Technologies
Slovak University of Technology Bratislava
Ilkovičova 3
842 16 Bratislava, Slovakia
jelemenska@fiit.stuba.sk

***Abstract*— The paper presents a knowledge assessment approach developed for support of Digital system description course. In the course the students should comprehend the methods and techniques used in digital system design, and gain the skills in digital systems modelling using hardware description languages (HDL). For many years the course exams used to be done only in pen and paper form. The students hated writing HDL programs on paper and the teachers disliked correcting the programs on paper as well. What is more, this type of exam has not allowed the verification of the student's ability to debug the created model, which is the substantial part of the student's practical skills. That is why, some years ago, we concentrated our work on designing a knowledge assessment system that will enable reliable evaluation of practical skills in the area of digital system description without substantial teacher involvement. Several applications have been developed and tested in the course. One of the best will be presented here.**

I. Introduction

Knowledge and skills assessment has always been problem in the courses devoted to programming languages, whether it was assembly or higher level programming languages [3,4]. The problem is very similar in the case of hardware description languages (HDL). To master the programming or modeling language, the students should solve various problems using the language. It is useless to learn language constructs or exact syntax by heart without writing any program. That is why the programming courses assessment process often adopt the following schema.

The lab exercises are based on more or less individually allocated assignments that the students can solve at school or at home. The solutions are evaluated by teachers and usually form a part of the overall course assessment. Unfortunately, there are always some students who try to find the easiest way to get the lab work results (e.g. search for a complete solution on the internet, copy the solutions from their classmates, etc.). The lab teachers usually have no chance to find out whether the students really did the programming work themselves, or not. They can just check up, whether the students understand their solutions and whether they are able to perform some minor changes in them. But it is not an easy task to do. Another problem is that different teachers might use various scales for assignments evaluation. As a result, many students do not get the required knowledge and practical skills, even though they get the highest evaluation for lab assignments. This part of the course assessment is often neither objective nor realistic although requiring a lot of teachers' work. The situation is even worse when distant learning is considered, where teachers have do not have the possibility to check out the students understanding of the solutions.

The second part of the course assessment is usually formed by an exam, sometime supplemented by some midterm tests. The tests and exams are often performed in a traditional way – using pen and paper. Although this type of exam gives minimum chance for cheating, there are many reasons why this type of exam is not suitable for practical skills assessment. First, the students have no chance to check the correctness of their programs or HDL models before handing them out for evaluation. Secondly, this type of exam does not allow verification of the student's ability to debug the created program/model, which is a substantial part of student's practical skills. Finally, hand written works are very difficult to read, correct, and evaluate. It is very demanding and time consuming work for teachers to evaluate programs/models written on paper and the result is often uncertain. In order to make the course assessment process more objective, fair and effective, there was an urgent need for redesign of the assessment. One of the solutions assumes that the substantial part of the course assessment will be shifted from the assignments to the skills based midterm tests. At the same time, new technology will be developed to perform the exam and midterm tests. The new applications, specially designed for this purpose, will enable the testing of not only the knowledge, but also the skills used, all in an automated manner.

II. Learning Management Systems Capabilities

To ease the work needed for knowledge and skills assessment in the course, various Content Management Systems (CMS), Learning Management Systems (LMS), Computer-Based Training (CBT) or Web-Based Training (WBT) systems can be used which support automated knowledge and/or assignments assessment. Nowadays there are a vast number of such systems, but most of them are unable to evaluate a student's skills using an online test.

Various systems supporting knowledge testing and/or assignment assessment have been reviewed in [2]. The emphasis was placed on supported questions, test types and automated assessment. The existing systems can be divided into two groups.

Test Creation Tools: These tools represent standalone solutions. The tests created using these tools can not usually be connected to learning content and must be used en bloc. The created tests conform to e-learning standards and therefore can be integrated to LMS. Another advantage is that these systems

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_22, © Springer Science+Business Media B.V. 2010

TABLE I.
ASSIGNMENT AND QUESTION TYPES SUPPORT IN REVIEWED SYSTEMS

	WebToTest	QUIZIT	Toolbook Instructor	eDoceo	Program Autor	Macromedia Authorware 7	ATutor	Dokeos	Ilias	HotPotatoes	Moodle
Assignment management	-	-	+	-	-	-	+	-	+	-	+
Question types											
Program	+	-	-	-	-	-	-	-	-	-	-
Multi-choice single answer	-	+	+	-	+	+	+	+	+	+	+
Multi-choice multi answer	-	+	+	-	+	+	+	+	+	+	+
True/False	-	+	+	-	+	+	+	-	+	+	+
Gap filling	-	+	+	-	+	+	+	+	+	+	+
Matching	-	+	+	-	-	-	-	+	+	+	-
Slider	-	-	+	-	-	-	-	-	-	-	-
Arrange objects	-	-	+	-	-	-	-	-	+	+	-
Place object	-	-	+	-	-	+	-	-	-	+	-
Identify correct picture region	-	-	-	-	+	+	-	-	+	-	-
Identify correct object	-	-	-	-	-	+	-	-	-	-	-
Crossword	-	-	-	-	-	-	-	-	-	+	-

+ =YES - = NO

are better elaborated and support more question types, and test creation is easier and flexible.

Learning Management Systems (LMS): These systems support e-learning and provide assignment management subsystems as well as testing subsystems. All considered LMS support the e-learning standards SCORM, ADL and IMS.

The capabilities of the reviewed systems are summarized in Table I. As we can see most of the reviewed systems support several types of test questions but only WebToTest enables students to edit, compile, and debug a program and submit it using a special form. We were not able to find any system supporting skills based assessment in the area of HDL modelling. This fact inspired us to develop several skills assessment systems supporting languages like VHDL and SystemC.

The standalone testing application presented in [5,12] and the Drag&Drop Moodle module [13] that have been used for student's knowledge and skills assessment in the last academic year are two examples of them. The special VHDL testing module, presented in this paper, is one of the latest solutions, suitable for online skills based examination in the area of digital systems modelling using VHDL. The module integrated into the Moodle LMS is already a part of the set of assessment tools used in the Digital system description course.

III. VHDL MOODLE MODULE DESIGN

A. Functional Requirements

There are three types of users in the module – student teacher and system.

1) Student: Student is the user with minimum rights. He or she can do only actions allowed by users with higher rights and therefore cannot modify functionality of the system.

Authentication – Students must log in to the system using the Moodle built in authentication system. They will have access to the Digital system description course, where a VHDL test is created as a special activity.

Source code editing – A test consists of test assignment and VHDL source code editing windows, which can contain a program core created by a teacher. Comfortable work during the test will be provided by VHDL syntax highlighting, similar to that used in ModelSim.

Source code compilation – Students can compile their source codes unlimited number of times. The compilation errors are returned to students by the system.

Source code simulation – The compiled model can then be simulated, using the test bench, prepared by the student or by the teacher. The student should receive the signals waveforms as an output.

Source code submission – The students have the possibility to upload their partial solutions to the server any time they wish. After the final submission, the system will further process the student's solution to get it evaluated automatically. The final submission of the source code is activated automatically in case the time allocated for the test elapses.

Results review – Students can see the results of all their VHDL tests in the separated module.

2) Teacher: The teacher can administrate his/her teaching module and adapt it to his/her own needs (i.e. adding teaching aids, activities, tests, creating forums etc.).

The VHDL test will be created adding the special new activity into the course. There are several activity attributes that can be set during this process.

Instance name – Using this name the teacher can locate and download or review the results of the respective activity.

General instructions – These instructions are optional and can be used to display any remarks made concerning the test.

Start of the test – This attribute can be used to schedule the test at any date and time. The test will not be accessible to students before this time.

Test duration – The teacher has to set up the time that the students can spend working on the test. After this time the final source code submission, followed by test evaluation is activated automatically.

Test assignment – The teacher creates a text form of assignment saying what the students should design.

Program core - The teacher can copy parts of the solution to the core which will be displayed in student's VHDL source code editing window. Typically at least the entity declaration of the designed circuit is provided to make the student's design consistent with the reference. Also, the teacher's test bench can be provided here, in case it should be made visible to students. Although this part is voluntary, if the teacher doesn't use it, the system will not usually be able to do automated evaluation.

Reference test bench – This test bench will be used to replace the student's one to make the final simulation after the final source code submission.

Reference solution – The complete solution including reference test bench should be provided here. This solution is used to generate the reference waveform that will be used for automated evaluation. The reference solution can also be used to make some source code comparisons.

Other parameters of the test include maximum number of points, difference penalty, syntactical correctness gain, and the figure to be displayed.

After the test, the teacher can use another module to review the solutions, eventually to alter their evaluation. There is also the possibility to download all the solutions of the selected test for further revisions or archiving.

3) System: Automated tasks are realized by the system. The rights must carefully be chosen to prevent system malfunction.

Source code storage – The system must guarantee the student's solution storage any time it is required to do so. The source code must be stored correctly even though it may contain special characters such as apostrophes, etc.

Source code compilation – Source code is compiled each time a student requires it. The actual compilation is done invoking the *vcom* compiler, an integrated part of ModelSim.

Source code simulation – The compiled model is simulated each time a student requires it. Simulation is done invoking the *vsim* simulator, an integrated part of ModelSim.

Final source code submission – This process is activated by the system either automatically in case the time allocated to the test elapses, or interactively by the student's command. As part of this process the student's test bench will be replaced by the teacher's, and the test will be automatically evaluated, comparing the resulting waveform to the reference. For each difference the assigned penalty is subtracted from maximum number of points. In case the source code can be compiled, but there are too many differences, a syntactical correctness gain is returned as a result. In case the source code is not syntactically correct, it can still be compared to the reference solution to find some partial similarities.

B. System Architecture

A client-server architecture model was used as shown in Fig.1. The LMS Moodle, together with the new module, are installed on the *server*. This web-based application requires a web server – in this case the Apache HTTP Server is used. A database server is also required. Although several types of databases can be used, the MySQL database is probably best supported. On the *client side,* a web browser is used to communicate with the server using a specially developed user interface.

The most important module functions are source code syntax verification and design simulation. To compile and simulate VHDL code, the commercial software ModelSim from Mentor Graphics Company is used. The compiler and the simulator can be launched from the command line and therefore initiated from another application. The *vcom* output - compilation errors - can be redirected into the file. The output of *vsim* is a .wlf format file, containing simulation results.

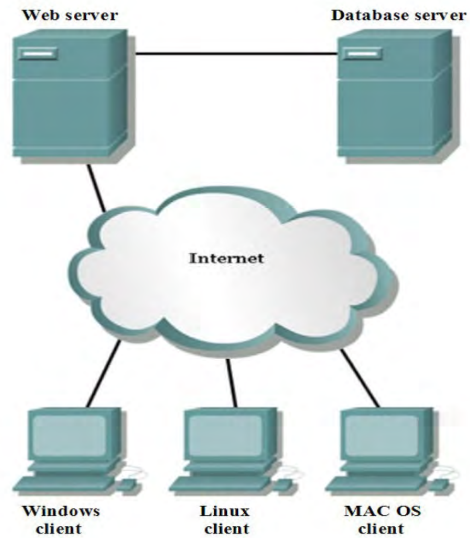


Fig.1. The system architecture

ModelSim is placed on the server. It is called by a PHP application using a console mode and provided source code as an input. ModelSim processes the provided input and returns the output to the PHP application. Using this solution the problem of server overloading might arise in the case of too many users. We can then try processing the requirements in parallel using, for example, a connected cluster of processors.

TABLE II.
TABLE MDL_VHDL: STORES DATA FOR MODULE INSTANCES

Name	Description
ID	Identifier of module instance
Course	Identifier of course where instance belongs
Name	Name of a course where instance belongs
Nadpis	Assessment title
Intro	Opening instructions and test directives
Zaciatok_testu	Time in epoch format when the test will start
Dlzka_testu	Test duration in minutes
Text_zadania	Text of assessment
Kostra_kodu	Source code scheme in VHDL
Testovacia_vzorka	Test sample which will be used for compilation and simulation
Obrazok	Name of the picture which will be used during test
Pocet_bodov	Maximum number of points
Bodova_zrazka	Points discount for one difference
Body_kompilacia	Base number of points which are given to student in the case of successful compilation

IV. VHDL Moodle Module Implementation

System implementation uses the MySQL database. This database is used by the Moodle system as a storage place for all data except files, stored on disc. The VHDL testing module uses data from this database as well. However, some new database tables had to be created to store data associated with this module.

A. *VHDL Testing Module*

Two new database tables were created in order to store data associated with this module: *mdl_vhdl* in Table II. stores data for module instances, and *mdl_vhdl_odovzdane* (shown in Table III.) stores the source code of particular students. The fields identify the specific test and student, as well as the full source code and results of the evaluation. The first names and surnames of students are stored for test archiving purposes – even if the Moodle user database is cleared, the complete tests are still available.

B. *Module for Reviewing Submitted Assignments*

This module also needs to have its own table, where all created instances are stored. The table is called *mdl_prehlad* and contains only 3 items: ID, Course, and Name.

No other tables are necessary since the module uses data from the *mdl_vhdl_odovzdane* table, which holds all of the vhdl code submitted by students.

C. *Shell Script for Compilation and Simulation*

To compile and simulate the VHDL model a shell script is called from PHP application. It consists of three main branches.

1) The compilation branch creates a new subdirectory in the directory /tmp/data/, named using user ID. In the subdirectory, the library work and file source.vhd are created. Thereafter, the file is compiled with the *vcom* compiler. The compiler output is returned to PHP application.

TABLE III.
TABLE MDL_VHDL_ODOVZDANE: STORES SUBMITTED SOLUTIONS

Name	Description
ID	Identifier of submission
Course	Identifier of course which was used for examination
Test	Test name which was submit
User_id	Identifier of user who submitted the test
Zadanie	Assignment of the test which was solved by user
Kod	Code which was sent by user
Cas_zacatia	Time in epoch format when the test was started
Cas_odovzd ania	Time in epoch format when the test was submitted
Hodnotenie	Points evaluation of student
Result	Results for compilation and simulation
Meno	First name of student.
Priezvisko	Student’s surname.

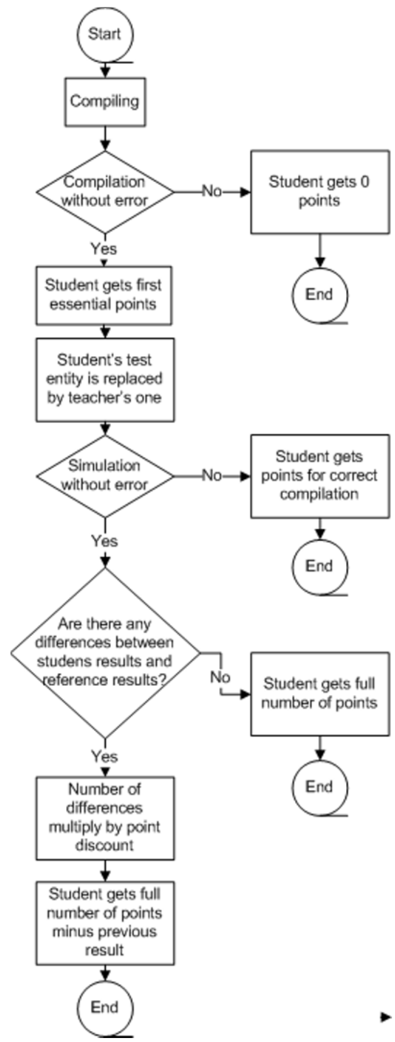


Fig.2. The correction algorithm

2) The simulation branch is a bit more complicated. At first, the existence of library work is checked. If it is there, the simulation *.do* script is dynamically created. This script is sent to the *vsim* simulator and causes the test bench simulation to be run. It is necessary to name the test bench “TESTBNCH”. If simulation runs without errors, the reference and student’s waveforms (simulation outputs) are compared. The comparison .do script is dynamically created for this purpose and sent to the *vsim* simulator. The outputs of the simulation and the comparison are returned to the application.

3) The final evaluation branch cleans up all the temporary files created during the test. Then the file source.vhd is created once more, using the reference test bench in this case. The file is compiled and simulated analogously. The outputs are then returned to the PHP application where they are processed and the results are calculated. The algorithm for the evaluation of student solutions is given in Fig. 2.

D. Test Creation Shell Script

After the instance of the testing activity is created by the teacher, a reference waveform in the form of .wlf file is automatically generated. A shell script is used for this purpose, which is called after the test creation form is sent. This script compiles and simulates the reference solution entered by the teacher. The output of the simulation, a .wlf file, is then stored together with the other files associated with this testing activity instance (like e.g. image). They are all stored in the directory *moodledata/courseid/instanceid*.

V. VHDL MOODLE MODULE USER INTERFACE

A. New Instance of VHDL Test Creation

The screenshot given in Fig. 3 represents a part of the test creation form. This form is only accessible to the teacher or administrator. The teacher sets all the test attributes here - for example, duration of test, test bench, reference solution, etc.

B. VHDL Test Elaboration

The screenshot shown in Fig. 4 illustrates the student interface to the designed module. Shown at the top of the screen are the remaining time, the test assignment text, and an optional scheme. Under these parts is the VHDL source code editing window, where the provided core should be completed using the student's VHDL code. VHDL syntax highlighting, similar to that used by ModelSim (used during the term), is used to make the editor familiar to students.

Fig. 3. The teacher screen – Test creation

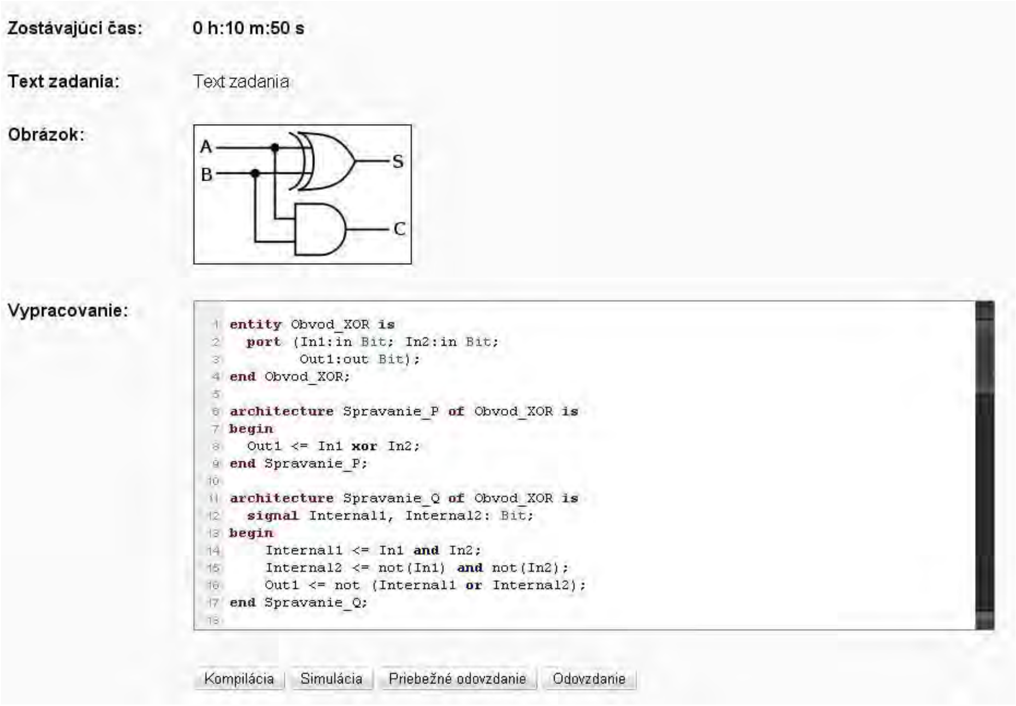


Fig.4. The student screen – Test elaboration

VI. RESULTS AND EXPERIENCE

We started to use the new testing module in the course Digital systems description to evaluate students' practical skills during midterm tests. In the course, the students will pass three midterm tests and one final exam. On each midterm test, the students are divided into three groups based on the testing environment - TabletPC, Drag&Drop, or VHDL Moodle test. Each student will pass each type of test once. However, the test assignment is the same, regardless of the test type. This constraint ensures that the severity of all the types of the test should be more or less the same. Fig. 5 shows the results of the first and the second midterm test - the distribution of Average Test Attainment. As we can see, the students reached the best attainment in the Drag&Drop test. The reason is that in this case the correct solution is already given in the test and the students just have to find the correct statements and place them correctly into the VHDL model.

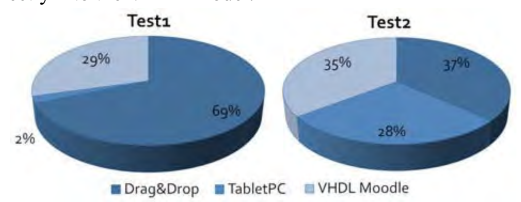


Fig. 5. The first and the second midterm test results

After the test students were asked several key questions. The answers to these questions are shown in Fig. 6. As we can see, most of the students think that the Drag&Drop test, composing a model from given statements, is the best way of testing practical skills.

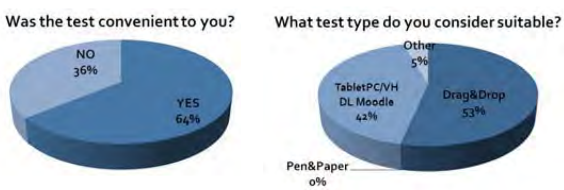


Fig. 6. Survey results

VII. CONCLUSION

The first results prove that the kind of online test described previously provides more realistic image of student knowledge of digital systems modelling, compared to tests based on the types of questions commonly used in online tests. Because the VHDL Moodle test requires creativity by students, the test results are worse compared to the Drag&Drop test.

The presented testing solution brings number of advantages. First, there is a substantial reduction in the demanding and time consuming work of the teacher, relative to paper exams correction. At the same time, the level of difficulty of the exam is preserved. Second, it enables the teacher to check the skill of students in the area of model debugging, which was previously not possible. All participants in the examination have at their disposal a source code editor with syntax highlighting, a VHDL language compiler, and simulator outputs. The application gives the teacher a means to comfortably manage the testing system. Finally, the students have the opportunity to verify their designs before sending them for evaluation.

ACKNOWLEDGMENT

The support by Slovak Science Grant Agency (VEGA 1/0649/09 "Security and reliability in distributed computer systems and mobile computer networks") and HP Technology for Teaching grant 2007 "Mobile Education Center" is gratefully acknowledged.

REFERENCES

[1] M. Mayer, "Distance Learning Support for Digital Systems Description Course," master theses, FIIT STU Bratislava, 2009.

[2] P. Polačko, "Learning Support for Specification and Description Languages Course," master theses, FIIT STU Bratislava, 2007, pp. 86

[3] Ch. Douce, D. Livingstone and J. Orwell, "Automatic Test-Based Assesssment of Programming," ACM Journal of Educational Resources in Computing,Vol. 5, No 2, September 2005

[4] C. A. Higgins, G. Gray, Pavlas Symeonidis and A. Tsintsifas, "Automated Assessment and Experience of Teaching Programming," ACM Journal of Educational Resources in Computing,Vol. 5, No 3, September 2005

[5] Š. Beleš, K. Jelemenská, M. Jurikovič, P. Pištek, T. Polák, J. Zeman and M. Žucha, "Mobile Testing System," Košice : Elfa, 2008, In: ICETA 2008 : 6th International Conference on Emerging e-Learning Technologies and Applications. High Tatras, Slovakia, September 11-13, 2008. - Košice : Elfa, 2008. - ISBN 978-80-8086-089-9. - pp. 133-138

[6] Š. Beleš, M. Jurikovič, P. Pištek, T. Polák, J. Zeman, and M. Žucha – "Learning Support for Specification and Description Languages," Team Project, FIIT STU Bratislava, 2008, pp. 83

[7] J. Kytka, "Learning Support for Specification and Description Languages Course," master theses, FIIT STU Bratislava, 2006, pp. 85

[8] N. Pollock, J. Cornford, "Theory and Practice of the Virtual University, report on UK universities use of new technologies," http://www.ariadne.ac.uk/, Jun-2000

[9] Bruggeman, J., Long, H., M., Misra, C.: Effective IT Security Practices (ID: EDU06165), Presented at EDUCAUSE Annual Conferences (10/09/2006),

[10] Brown, M., B., Smith, J., M., Windham, C.: From Today's CMS to Tomorrow's Learning Support System, Presented at EDUCAUSE Annual Conferences (10/11/2006), WINNER: 2006 EDUCAUSE Catalyst Award.

[11] Hemmi, A., Pollock, N., Schwarz, C.: If not the Virtual University then what? Co-producing E-learning and Configuring its Users, 8th European Conference on Media in Higher Education, (Best Overall Paper), GMW, University of Duisburg, 16-19 September 2003.

[12] K. Jelemenská, M. Jurikovič, P. Pištek and P. Čičák, "Automated Skills Assessment in Mobile Environment," in: Research, Reflections and Innovations in Integrating ICT in Education, Vol.1, V.International Conference on Multimedia and Information & Communication Technologies in Education (m-ICTE 2009), Lisbon, Portugal : Vol. 1. - Badajoz : Formatex, 2009. - ISBN 978-84-692-1789-4. - pp. 487-491

[13] K. Jelemenská, E. Tomalová, and P. Čičák, Pavel, "Knowledge Assessment Improvement in Moodle Environment," in: Research, Reflections and Innovations in Integrating ICT in Education, Vol. 2, V.International Conference on Multimedia and Information & Communications Technologies in Education (m-ICTE 2009, Lisbon, Portugal. - Badajoz : Formatex, 2009. - ISBN 978-84-692-1790-0. - pp. 979-983

[14] J. Genči, "Knowledge Assessment – Practical Example in Testing," In: Journal of Communication and Computer, Vol. 6, 8, 2009, pp. 65-69, ISSN 1548-7709

Wikibooks as Tools for Promoting Constructivist Learning in Higher Education: Findings from a Case Study

Ilias Karasavvidis
Department of Preschool Education
School of the Humanities
University of Thessaly
ikaras@uth.gr

Abstract - **The present paper focuses on wiki-related pedagogies in higher education and examines the potential of wikibooks in a blended learning context. A wikibook was integrated as a compulsory class assignment in an undergraduate course in a preschool education department. Twenty seven students participated in the study which investigated (a) the extent of student participation in the wikibook and (b) students' perspectives on the contribution of the wikibook to their learning. In stark contrast to findings reported by other studies, results indicated that the rate of student participation was very high as suggested by indices such as page views, page edits, and number of students involved per page edit. Moreover, the students rated positively wikibooks' contribution to the learning of the course concepts and the active mode of learning required. However, the students also mentioned significant problems with the wikibook activity such as difficulty understanding certain sections and concerns with the validity of their writings. The paper is concluded with a discussion of the findings and implications for further research.**

Keywords – Wiki, Wikibook, Web 2.0, Collaborative Learning

I. Introduction

A. The second generation of the world wide web

Web 2.0 has attracted considerable attention over the past few years. Currently, the emphasis is on collective intelligence, the wisdom of the crowds, wikinomics, distributed asynchronous collaboration, social networks and networking, social software, massively multiplayer on-line games, and virtual worlds. According to [1], Web 2.0 can be defined in a number of ways. On a technical level, Web 2.0 tools such as blogs, wikis, podcasts, social bookmarking, photo sharing, instant messaging, VoIP applications, RSS feeds, mashups etc afford more possibilities for communication and collaboration compared to Web 1.0 tools. For example, wikis have considerable advantages over threaded discussions [2] and blogs have more affordances compared to Content Management Systems (CMS). Regardless of technical differences, however, what sets apart Web 2.0 from the first generation of the Web is the role users play.

In the first generation of the world wide web users were simply browsing information which was provided on web pages. In the case of Web 2.0 the users are not simply consumers of information available on web sites. Instead, given the architecture of web 2.0, the users are actively contributing all sorts of content. In fact, most of the typical web 2.0 tools do not only rely on user participation but would actually be inconceivable without it. For example, web sites such as Wikipedia and YouTube would not exist without user-generated articles and videos respectively. Consequently, the first generation of the web, now called Web 1.0, has been transformed into a new, participatory, read-write web which is known as Web 2.0. The most distinctive feature of Web 2.0 is that it relies on the wisdom of crowds [3] which in turn leads to crowdsourcing and wikinomics [4].

B. Web 2.0 and E-Learning 2.0

The impact of Web 2.0 technologies has been substantial [5]. E-learning is one of the areas of education where Web 2.0 has had a major impact [6]. Terms such as Education 2.0, School 2.0, Classroom 2.0, and Learning 2.0 are becoming commonplace.

The importance of Web 2.0 for enhancing learning in higher education is being increasingly recognized. Its significance lies in that it can address some of the problems which characterized several E-learning implementations. One of the most often voiced criticisms is related to the adopted e-learning pedagogy. More specifically, it has been argued that many E-learning environments are content-driven and are being mostly used to deliver course materials on-line ([7], [8], [9]). Such transmissionist approaches, which are dominant in E-learning, lead to the replication of traditional modes of teaching and learning on-line [10].

As opposed to traditional E-learning, the role of students in the new E-learning 2.0 era is more active and participatory. The students are not consumers of material which has been compiled by instructors ([9], [5]). One of the major tenets of Constructivism, the current dominant learning paradigm in higher education, is that learning can only be achieved through the active student participation in the learning process ([11], [12], [13]). This active participation by learners is perfectly compatible with the notion of active participation by users in the context of Web 2.0. Thus, the affordances of Web 2.0 technologies are in line with the active approach to learning which constructivism requires, which constitutes the added value of E-learning 2.0 from a learning perspective. Web 2.0 technologies can help transform traditional higher education practices which are based on objectivist notions of learning and are characterized by learning metaphors that are transmisionist in nature.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_23, © Springer Science+Business Media B.V. 2010

C. Wikis

Much like all other Web 2.0 tools, wikis have attracted widespread attention. "WikiWiki" is a Hawaiian word that means quick. According to [14], a wiki is a "*freely expandable collection of interlinked webpages, a hypertext system for storing and modifying information – a database, where each page is easily edited by any user with a forms-capable Web browser client*" (p. 14). Essentially, a wiki system constitutes a collection of webpages with information which can, in principle, be edited by anyone. Users can add content to existing pages, restructure the information on existing pages as well as create new pages and create links between existing and new pages. The fact that everyone can create, edit and delete a page means that wikis have a very open-ended nature. Cunningham developed the first wiki, WikiWikiWeb, in 1995 and there are more than 200 clones today.

The popularity of wikis is due to their simple and flexible nature. Firstly, wikis are characterized by a non-linear hypertext structure: a wiki enables the creation of associative hypertexts where each page contains links to other pages and the user can choose which links to follow. Secondly, wikis are characterized by ease of access since all a user needs to use a wiki is a web browser. As a web application the wiki does not require any specific browser or operating system software. Thirdly, the technical skills required for participation are minimal as users do not have to learn HTML to create a webpage. Thanks to wiki markup, a simplified version of HTML, and the built-in WYSIWYG editors, the burden for creating and formatting a webpage is significantly reduced. Finally, a wiki system also provides other functions such as (a) tracking of edits through a logging and notification system, (b) comparison between different versions of edits, (c) rollbacks to earlier versions, (d) discussions per wiki page and (e) customizable read and edit access rights as well as protected pages.

Wikipedia, the prime example of wiki success, constitutes an indication of how popular wikis have become.

D. Wikis in Education

Wikis can be used to support a wide range of activities including but not limited to documentation, reporting, project management, on-line glossaries and dictionaries, discussion groups, and information systems. Wikis can be used to promote learning, collaboration, communication, interaction, sharing, meaning making, and reflection. [15] outline three main types of uses for wiki:

(a) distributing information: the wiki becomes the medium to disseminate information as well as the place to store the information

(b) collaborative artifact creation: text is the main artifact produced, be it in the form of glossary or other formats

(c) discussion and review activities: anchored discussion, project case libraries, and peer review.

Even though wikis can be very important for education, there is a knowledge gap in the literature. Research on wiki applications shows that wikis can be very useful tools for supporting learning in higher education settings. More specifically, it has been reported that wikis can facilitate group work [16], collaboration [17], collaborative learning [18], inquiry learning and co-reflection [19], and peer assessment [20]. Regarding performance gains, there is some evidence to suggest that the use of wikis can lead to improved course performance (see e.g. [21], [22]) and improved writing and critical thinking skills [23]. Finally, students appear to have positive attitudes towards wiki use ([24], [25]) and find their wiki experiences to be valuable for their learning ([22]; [17]).

Despite positive findings, significant challenges have been identified: (a) minimal student participation ([26], [27], [28], [21]), (b) limited on-line interaction ([23]; [29], [30], [22]), (c) preference for individual work [17], and (d) negative student perceptions of wiki's contribution to their learning ([26]; [30]; [22]). While this brief literature review does suggest that wikis hold great promise for higher education, one of the critical issues which remain to be addressed is related to pedagogy. There is a knowledge gap regarding how wikis can be best incorporated into the teaching and learning practices of higher education. The evidence from existing research clearly suggests that the successful wiki integration into a course requires mildly put a thorough redesign ([21], [28]).

E. Focus of the study

The present study aimed to fill this knowledge gap in wiki-related pedagogies in higher education. The specific focus in this study is wikibooks, one of the many possible uses of wikis in university courses.

Wikibooks can be considered as class assignments whose main objective is to engage students in the collective writing of an electronic textbook. This effectively combines the active approach to learning which is required by constructivism and the unique potential of wikis to be used both (a) as process tools to facilitate collaboration and (b) product tools to enable the creation of a shared artifact such as a textbook.

The potential of wikibooks has been researched in a number of recent studies in higher education settings (see e.g. [31], [22], [32], [33], [34]). The typical use of wikis in those studies involves the writing of textbooks by students as part of the standard course assignments. As a rule, the findings reported by those studies are favourable in terms of academic performance, level of student engagement, and student attitudes. It should be stressed, however, that the aforementioned studies are either case studies or quasi experimental ones and, as such, are mostly exploratory. Consequently the evidence is far from conclusive. In fact, despite positive findings significant problems also emerged. More specifically, some students in the [22] study argued that they did not learn anything new from the wikibook activity because they only wrote what they already knew. Along the same lines, [34] argued that it is naïve to anticipate that the wikibook activity would necessarily lead to collaborative writing and learning without taking any other measures.

The present study examines the potential of wikibooks in tertiary education in a blended learning context. A wikibook activity was incorporated in an undergraduate course in a preschool education department as part of the typical course assignments. Given the wikibook use in such a context, the study aimed to address the following research questions:

(a) *what is the rate of student participation in the wikibook?*

(b) *what are students' perceptions of their experiences with the wikibook?*

II. Method

A. Subjects

27 students participated in the study. The students enrolled in an undergraduate course on the educational applications of the Internet which was taught by the author in his parent institution. The course was elective for fourth semester students and this cohort of students constituted a convenience sample.

B. Course

The course aimed to introduce (a) common internet technologies and services and (b) a set of pedagogical principles which could support educational uses of the internet. The course comprised two main components: technical and pedagogical. The main objective of the technical component was to introduce students to information, communication, file, and application services. The main objective of the pedagogical component was to use the internet services to support a broad range of teaching and learning activities.

Given that wikis were among the Web 2.0 tools introduced in the context of the technical component and that the new learning approaches were also presented in the context of the pedagogical principles, the wikibook afforded a unique opportunity to put theory into practice. The students could not only familiarize themselves with the wiki technology and the associated pedagogical ideas but also experience it firsthand.

F. The Wikibook assignment

The wikibook constituted the main course assignment which was compulsory for all students. This assignment required students to (a) gather information on a number of topics from various sources (e.g. printed textbooks, electronic textbooks, wikibooks, web sites etc), (b) process the information (e.g. restructure, summarize, compare, contrast etc), and (c) write the information to a wiki page thereby contributing to an existing chapter or sub-chapter.

Regarding participation, no specific guidelines were provided: the students could freely decide on how much information to contribute to the wikibook. Moreover, there were no specific requirements set for covering a minimal number of topics or chapters: the students could contribute information to topics which they found interesting. To encourage participation and avoid social loafing, the wikibook assignment amounted up to 60% of the total course credit.

MediaWiki (http://www.mediawiki.org), the software that powers Wikipedia, was the software used in the study. To avoid spam, vandalism, and all sorts of extraneous distractions, the wikibook was used in a closed-study group.

G. Measures

For the purposes of the present study both quantitative and qualitative measures were employed. Quantitative measures included the wiki system log files and two survey instruments. Using the wiki log files, a number of measures were computed: total number of pages created, number of page views, number of user views per page, number of edits, and the number of users involved per page edit. The two survey instruments were developed ad hoc for the purposes of the study. The first instrument measured student background knowledge and attitudes towards ICT. The second instrument measured student views about their wikibook experience. It contained 15 statements on several aspects of the wikibook and a typical 5-point Likert scale was used. Qualitative data collection involved 8 semi-structured interviews which aimed to provide a richer understanding of how students viewed the wikibook.

H. Procedure

The students were introduced to the wiki at the beginning of the course. All students received an extensive introduction in using the wiki: in the sandbox provided students experimented with all technical aspects of the wiki such as creating new pages, revising existing ones, linking, rolling back, and discussing. After each class lecture or lab session the students selected a set of concepts which they deemed both important and interesting, searched and gathered relevant information, wrote up the text and contributed the information to the wikibook.

In the beginning of the course the students were surveyed regarding background knowledge of and attitudes towards ICT. At the end of the course the students were surveyed about their learning experiences with the wiki and eight students were randomly selected for interviews.

III. Analysis

To address the first research question, data analysis involved frequency counts and descriptive statistics. To address the second research question descriptive statistics indices were computed for the two surveys while the interviews were content-analyzed to determine the main themes the students reported.

IV. Results

A. Rate of student participation

The students created 66 pages which were viewed 6909 times. The contents page (i.e. the wiki start page which is the equivalent of the table of contents) is given in table 1.

Internet
World Wide Web
The social impact of the Internet
File Sharing
Secure Data Transfer
Networks
Distance Education
Hacking & Cracking
Portals
Operating Systems
Important Internet Applications
Internet & Schools

Table 1: The chapters of the wikibook

The main wiki statistics are presented in table 2. As can be seen in the table, some pages were more popular than others. In fact, the wiki contents page amounted to 88% of the total number of views. On average, each page was read 104 times and the variability was very high. Log file analysis, however, indicated that 55 out of the total 66 pages were created by one student and were characterized by a small number of views (typically 1-3). While technically these constituted wiki pages, they contained little information -usually an image or icon and a couple of sentences describing the theme of the page. As this student case clearly represents an outlier, the descriptive statistics reported in table 2 should be interpreted with caution. To address this issue, we examined the page views for the 13 most viewed pages, that is, the contents page (main wiki entry page) and the 12 other pages listed in that page. This subset of wiki pages accounted for 88,57% of the total page views and did not include any of the wiki pages created by the aforementioned student. After this correction, the mean number of reads increased by a factor of five (M=470,69, SD=578,97). This finding clearly suggests a very high rate of access.

The wiki pages were edited 1065 times. Again, the statistics reported in table 2 are subject to the same distortion considering that most of the 55 pages that one student created had not been edited at all. Following the same procedure adopted above to deal with this outlier the descriptive statistics were recalculated. The contents page and the pages listed as the main wikibook contents were edited 757 times (71,08% of the total edits recorded by the system). The average number of edits per page was 58,23 which is considered to be very high and indicates that all pages underwent extensive revisions. It should be noted, however, that even for this subset of wiki pages the variability was rather high (SD=75,20).

The number of users involved per page edit is an important measure as it shows how many students were involved in the authoring of a wiki page. Table 2 shows that 2,7 students were involved per page edit. However, as discussed above, this figure is heavily biased because 83% of the pages were created by a single student. After correcting for this outlier, the mean number of students involved per page edit rises to 8,69 (SD=6,21). This suggests that the wiki pages were essentially co-authored as on average more than 8 students were involved in the writing of every wiki page. This finding is very promising as the number of joint edits reflects the degree of on-line student collaboration. Finally, in addition to obtaining evidence of on-line collaboration among students, table 2 also shows that the levels of individual student participation were also very high: the average number of edits per user is about 40 which suggests that students' contributions were extensive.

Variable	Total	min	max	Mean	SD
Page Views	6909	1	2045	104,68	308,8
Page Edits	1065	1	262	14,3	39,1
Users per Page Edit	n.a.	1	22	2,77	4,03
Edits per User	1065	3	184	39,44	44,94

Table 2: Descriptive statistics for the main Wikibook variables

For the most part the pages were reasonably interlinked. The contents page (initial wiki page) included links to 12 wiki pages through which the majority of all other pages were accessible. Only 9 pages (13,64%) were not linked to any other pages in the wiki.

B. Student perceptions

The student responses to the questionnaire administered at the end of the course are given in table 3.

	Variable	Mean	SD
general	ease of use [a]	3,85	0,66
	ease of access [a]	4,07	0,92
	overall impression [b]	4,22	0,58
	understanding contribution [c]	3,89	0,58
contribution	writing contribution [d]	4,19	0,40
	reading contribution [d]	3,81	0,62
	edits made contribution [d]	3,59	0,64
	edits received contribution [d]	3,59	0,80
	discussion contribution [d]	3,59	0,80
edits	extent of revisions made [e]	1,85	0,91
	extent of revisions received [e]	1,74	0,86

a. Very difficult-Very easy b. Very negative-Very positive. c. Not at all important – Very important. d. Strongly disagree-Strongly agree. e. Minor-Major

Table 3: Descriptive statistics of students' views regarding the wikibook

As table 3 shows, the students did not experience any difficulties in using or accessing the wiki and rated their overall wiki experience very positively (4,22). In fact, the mean for this variable reflected the highest positive rating in the questionnaire. Moreover, the students evaluated positively wiki's contribution to their understanding of the course concepts. While students rated positively the four aspects of wiki work, the contribution of writing to learning clearly stands out. This is quite interesting as the students appeared to favor writing over reading, editing, and discussing. Finally, table 3 shows that students reported that they neither made extensive edits to the writings of other students nor were their texts extensively edited by fellow students.

The students were also asked to rate the most difficult aspect of wiki-related work. The final rating is as follows: writing: 33%, reading: 11%, revising: 44%, discussing: 11%. Thus, revising was found to be the most difficult aspect of the wikibook assignment.

The qualitative analysis of the interviews indicated that while students expressed positive views about the wiki, they also mentioned considerable difficulties. An outline of the pros and cons is presented in table 4.

Pros	Cons
Facilitation of conceptual understanding the wiki task helped students master factual information	**first to write** as the first contributors usually summarized the main points presented in class, the remaining students reported difficulties in finding what to write
Information gathering & exchange the wiki task necessitated information gathering for the students	**extent of contributions** because no specific guidelines were given, the students were unable to decide how much information in each book section was enough
group work the wiki task required a collaborative effort	**structuring** the students considered the openness of the wiki as somewhat limiting and requested more structure (e.g. providing a detailed table of contents)
revising the task afforded the opportunity to revise not only one's texts but also the texts contributed by others	**understanding** the students complained that some sections of the wikibook were unreadable which obfuscated understanding
writing the students considered the wiki task to be a genuine text-production process and not a simple reproductive one	**validity of interpretations** the students were uncertain about the validity of the information which they added to the wikibook
mode of learning the wiki task exclusively relied on student contributions and this promoted a more active approach to learning in the context of the course	**technical issues** some students mentioned that they were sometimes unable to commit the edits they had made to a wiki page (probably due to concurrent edits)

Table 4: Students' perceptions of the pros and cons of the wikibook activity

V. Discussion

The impact of Web 2.0 on many domains of human activity has been substantial [5]. One of the areas where this impact is most evident is education and more particularly E-learning [6]. Considering that many web-based education and e-learning approaches have tended to replicate traditional modes of teaching and learning which are transmisionist in nature, Web 2.0 is expected to enhance learning in higher education because, by default, it requires the active participation of users. Unlike Web 1.0 tools, the corresponding Web 2.0 tools would not be at all possible without user participation. This specific feature is a perfect match to the active learner participation which according to Constructivism characterizes all human learning. Thus, Web 2.0 holds great promise for promotive active learning in higher education. Wikis are amongst the Web 2.0 tools which have great potential for contributing to learning because they are uniquely suited to support both process and product approaches. There are many types of wiki uses in education and the specific focus in this paper was wikibooks. Wikibooks can be used as class assignments where the students collaborate on-line and create an electronic textbook. A number of recent studies have began to explore how wikibooks can be used in the context of higher education ([31], [22], [32], [33], [34]). Although the findings from those exploratory studies are rather positive, a number of important issues related to pedagogy remain unexplored. The present study aimed to fill this knowledge gap regarding wikibook applications in blended learning contexts. The study examined both the extent of student participation in the wikibook activity and students' perceptions of the wikibook activity.

Regarding the first research question, the findings of the study indicate that the students created several pages which were both viewed and edited several times. As the analysis indicated, each wiki page was viewed on average 470 times. Moreover, each page underwent extensive edits as it was edited on average 58 times. Finally, more than 8 students were on average involved in the editing per wiki page. As a whole, these findings are very promising as most wiki-related studies typically report that student participation in the wikis is limited ([26], [27], [28], [21]) and the same holds for on-line interaction ([23], [29], [30], [22]). Thus, unlike former studies, the outcomes of the present one clearly suggest that a compulsory wikibook activity for which students can get a considerable portion of the total course credit can effectively alleviate the problem of limited student participation which is reported in the literature.

Regarding the second research question, the quantitative analysis indicated that the students rated their wikibook experience very positively as the mean of 4,22 on a 5 point scale suggests. As for the contribution of wikibook to their learning of the course concepts, the students gave higher ratings to writing and reading than to the remaining aspects.

The qualitative analysis of the student interviews revealed a number of positive and negative points of the wikibook activity. On the one hand, the students favored the wikibook activity for a number of reasons. It should be noted that all aspects of wikibook which were viewed as favorable by the students are in line with the theory and research regarding the potential contribution of wikis to learning. As far as the contribution of the wikibook to the learning of the course concepts is concerned, the facilitation of conceptual understanding and the active mode of learning that the wikibook required stand out from a learning point of view. On the other hand, the students also expressed considerable difficulties. Two of the problems students reported, namely understanding and validity of interpretations, deserve closer further examination because they are related to learning and, to a large extent, contradict the facilitation of conceptual understanding that the students mentioned as one of the most important wiki contributions. The remaining difficulties identified are more procedural in nature and can be more easily addressed (e.g. devising specific rules for wiki contributions or providing more structured versions of the task such as a detailed list of contents with chapters and sub-chapters). To a large extent, the outcomes of the qualitative analysis replicate other literature findings regarding the difficulties students experience when using wikis (e.g. [26], [30], [22], [35]). Thus, the fact that students face important challenges when engaged in wiki activities remains and should be addressed in future research.

References

[1] O'Reilly, T. (2005). *What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software*. Retrieved March 20, 2009 from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[2] West, J.A. & West, M.L. (2009). *Using wikis for on-line collaboration: the power of the read-write web*. San Fransisco: Jossey-Bass.

[3] Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Anchor.

[4] Tapscott, D. & Williams, A.D. (2006). *Wikinomics: how mass collaboration changes everything*. NY: Penguin.

[5] Mason. R. & Rennie, F. (2008). *E-learning and social networking handbook: resources for higher education*. London: Routledge.

[6] Downes, S. (2005). E-learning 2.0. *eLearn Magazine*. Association for Computing Machinery, Inc. Retrieved March 20, 2009, from http://elearnmag.org/subpage.cfm?section=articles&article=29-1.

[7] Naidu, S. (2003). Designing Instruction for e-Learning Environments. In M. G. Moore & W.G. Anderson (Eds). *Handbook of Distance Education* (pp. 349-365). Mahwah, New Jersey: LEA.

[8] Littlejohn, A., Cook, J., Campbell, L., Sclater, N., Currier, S. & Davis, H. (2007). Managing educational recourses. In G. Conole & M. Oliver (Eds). *Contemporary perspectives in e-learning research: themes, methods and impact on practice* (pp. 134-146). London: Routledge.

[9] Palloff, R.M. & Pratt, K. (2007). *Building on-line learning communities: effective strategies for the virtual classroom*. San Fransisco: John Wiley & sons.

[10] Laurillard, D. (2007). Preface. In H. Beetham, & R. Sharpe (Eds.), *Rethinking pedagogy for a digital age: designing and delivering e-learning*. London: Routlege.

[11] Jonassen, D. (1991). Objectivism vs. Constructivism. *Educational Technology Research and Development*, 39(3), 5-1.

[12] Jonassen, D. (1999). Designing constructivist learning environments. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory* (Vol. 2, pp. 215-239). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

[13] Jonassen, D., Cernusca, D., & Ionas, G. (2007). Constructivism and instructional design: The emergence of the learning sciences and design research. In R. A. Reiser & J. V. Dempsey (Eds.), *Trends and issues in instructional design and technology* (2nd ed., pp. 45-52). Upper Saddle River, NJ: Pearson.

[14] Leuf, B., & Cunningham, W. (2001). *The Wiki Way*. Addison-Wesley Professional.

[15] Guzdial, M. Rick, J. & Kehoe, C. (2001). Beyond Adoption to Invention: Teacher-Created Collaborative Activities in Higher Education. *The Journal of the Learning Sciences*, 10(3), 265–279.

[16] Choy, S.O. & Ng, K.C. (2007). Implementing wiki software for supplementing on-line learning. *Australasian Journal of Educational Technology*, 23(2), 209-226.

[17] Elgort, I., Smith, A.G. & Toland, J. (2008). Is wiki an effective platform for group course work? *Australasian Journal of Educational Technology*, 24(2), 195-210.

[18] Minocha, S. & Thomas, P. G. (2007). Collaborative Learning in a Wiki Environment: Experiences from a software engineering course. *New Review of Hypermedia and Multimedia,* 13(2), 187-209.

[19] Yukawa, J. (2006) Co-Reflection in on-line Learning: Collaborative Critical Thinking as Narrative. *International Journal of Computer-Supported Collaborative Learning*, 1(2), 203-228.

[20] Xiao, Y. & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *Internet and Higher Education*, 11(3-4), 186-193.

[21] Rick, J. & Guzdial, M. (2006). Situating CoWeb: a scholarship of application. *Computer-Supported Collaborative Learning*, 1(1), 89–115.

[22] Ravid, G., Kalman, Y.M. & Rafaeli, S. (2008). Wikibooks in higher education: Empowerment through on-line distributed collaboration. *Computers in Human Behavior*, 24(5), 1913–1928.

[23] Wheeler, S. Yeomans, P. & Wheeler, D. (2008). The good, the bad and the wiki: Evaluating student-generated content for collaborative learning. *British Journal of Educational Technology, 39*(6), 987–995.

[24] Robertson, I. (2008). Learners' attitudes to wiki technology in problem based, blended learning for vocational teacher education. *Australasian Journal of Educational Technology*, 24(4), 425-441.

[25] Theng, Y-L, Li, Y., Lim, E-P., Wang, Z., Goh, D., Chang, C-H., Chatterjea, K., & Zhang, J. (2006). Understanding User Perceptions on Usefulness and Usability of an Integrated Wiki-G-Portal. *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL)* (pp. 27-30). Kyoto, Japan.

[26] Carr, T. Morrison, A., Cox, G., & Deacon, A. (2007). Weathering wikis: Net-based learning meets political science in a South African university. *Computers and Composition*, 24(3), 266–284.

[27] Choy, S.O. & Ng, K.C. (2007). Implementing wiki software for supplementing on-line learning. *Australasian Journal of Educational Technology, 23*(2), 209-226.

[28] Cole, M. (2008). Using Wiki technology to support student engagement: Lessons from the trenches. *Computers & Education,* 52(1), 141-146 .

[29] Ma, W.W-K. & Yuen, A.H-K. (2007). Learning News Writing Using Emergent Collaborative Writing Technology Wiki. In J. Fong, F. L. Wang (Eds). *Blended Learning. Workshop on Blended Learning* (pp. 303-314): Edinburgh, United Kingdom: Pearson.

[30] Ma, W.W.K. & Yuen, A.H.K. (2008). A Qualitative Analysis on Collaborative Learning Experience of Student Journalists Using Wiki. In J. Fong, Reggie, K. Wang, Fu Lee, W.F. (Eds.). *Hybrid Learning and Education. Proceedings of the first International Conference on Hybrid Learning* (pp. 103-114). Lecture Notes in Computer Science 5169. Berlin: Springer-Verlag.

[31] Bold, M. (2006). Use of wikis in graduate course work. [Electronic version]. *Journal of Interactive Learning Research*, 17(1), 5-14.

[32] Kidd, J., O'Shea, P., Baker, P., Kaufman, J. & Allen, D. (2008). Student-authored Wikibooks: Textbooks of the Future?. In K. McFerrin et al. (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference* (pp. 2644-2647). Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/p/27619.

[33] Lai, F. & Chang, C. (2008). Wikibook Writing for Collaborative Learning. In G. Richards (Ed.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 977-982). Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/p/29735.

[34] Lin, H. & Kelsey, K.D. (2009). Building a Networked Environment in Wikis: The Evolving Phases of Collaborative Learning in a Wikibook Project. *Journal of Educational Computing Research*, 40(2), 145 – 169.

[35] Karasavvidis, I. (in press). Wiki uses in higher education: exploring barriers to sucessful implementation. *Interactive Learning Environments*.

The Impact of the Context of an Information System on Strategic Information Quality - The Institute of Technology Sector in Ireland

Laurence P. Elwood* and Owen V. Foley**

*School of Business, Galway Mayo Institute of Technology, Ireland. Email: larry.elwood@gmit.ie

**School of Business, Galway Mayo Institute of Technology, Ireland. Email: owen.foley@gmit.ie

***Abstract*— Information Quality has become an ever increasing problem for many organizations relying on information systems for many of their core business activities. Despite advancements in technology and information quality investment the problem continues to grow. The context of the information systems over the last number of years has also evolved. This has resulted in a diverse number of architectures accessing multiple information systems. The context of the deployment of an information system in a complex environment and its associated information quality problems has yet to be fully examined by researchers. Our research addresses this shortfall by examining the relationship between the context of the information system and information quality. The research is conducted within the context of strategic information quality within the higher education system in Ireland specifically the Institute of Technology Sector. A mixed methodology approach is employed to tackle the research question of the impact of the context of an information system on strategic information quality. Research to date has examined different information systems' context factors and their effect upon information quality dimensions both objectively and subjectively. Thus far our research puts forward a novel information quality methodology that is context related; that is it takes the user, task and environment into account. An initial experiment has been conducted. Results of preliminary experiments indicate that context affects the perception of information quality.**

***Keywords* — Information Quality; Higher Education; Strategy; Information Systems**

I. Introduction

Over the last two decades Information Systems (IS) contexts have radically altered [1-3]. The relatively simple mainframe – workstation configuration has evolved and is accompanied or replaced in its entirety. Contexts have become more complex. A typical IS is most likely accessed via a number of different tiers, applications and platforms. Therefore the context has changed from a homogeneous one of a single database and access method to a heterogeneous and complex one resulting in many databases and access methods. This landscape will continue to evolve with the ever increasing deployment of Service Oriented Architectures (SOA) [4].

The changing context of the IS comprises a number of factors; including the user role, application tasks, access devices and associated services. Lee [5] describes this as the relationship between contents and environments.

The workstation model was predominant in the late 70s and early 80s [2], followed by widespread deployment of the client server model [3]. The rapid expansion of internet services soon followed [1, 3]. This expansion has continued with the pervasive availability of mobile devices and wireless technologies [1, 6, 7]. Virtualization has also made an impact especially with the widespread implementation of unplanned solutions [8].

The significant changes and increase in complexity of IS context has in many cases occurred independently of the underlying databases accessed [1]. Applications may have been designed, built and tested with a mature software development methodology for a particular context, yet within a very short period of time it may be accessed and predominantly employed from a different context [9]. Data models have also evolved [2], nonetheless a considerable number of IS in use today have data models designed prior to the contexts that are employed to access them. There are as a result multiple accesses from diverse and complex contexts.

The quality of information that strategic decision makers require is heavily dependent on these IS. An examination of the Institute of Technology Sector in Ireland reveals that Management Information Systems, Decision Support Systems and User driven Office Information Systems are the primary source of this information. In light of the evolving nature of both the technology and the higher education sector it is imperative to measure the impact of these contexts on the strategic information quality (IQ).

II. Higher Education in Ireland

Ireland's higher education providers can be characterized across a university and institute of technology divide. The principal differentiator across both sectors is the applied orientation of the institute of technology sector, which may be likened to the fachhochschulen in Germany. The Institute of Technology sector was established in 1969 (when colleges were referred to as Regional Technical Colleges) following the earlier recommendations of the "Mulcahy Report" [10]. Prior to the Regional Technical Colleges Act [11], these colleges had limited autonomy and were controlled by the local education authority, the Vocational Education Committee. The Regional Technical Colleges Act [11] expanded the mission of the Regional Colleges permitting them for the first time to engage in research, participate in collaborative research and the commercial exploitation of

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_24,

research. Teaching nonetheless continues to be seen as the primary mission of the institute of technology sector, with research as an interesting, but ultimately costly, adjunct. The Regional Technical Colleges Act [11] also established the following bodies; Governing Body, Academic Council and the designation attaching to the principal officer of each institute, namely Director (re-designated to that of Institute President under the Institutes of Technology Act [12]. Consequently, three distinct groupings now exist in each college: (1) The Governing Body, which oversees the work of the College Director, (2) the College's senior management team, headed by the College Director, which oversees the day to day management of the College, and (3) the College's Academic Council, charged with protecting, maintaining and developing the academic standards of the courses and activities of the college. Two of these bodies; Governing Body and the senior management team, are essentially managerial in focus, requiring data which facilitates the discharge of their managerial duties. The mid 1990s saw further pressure on Ireland's higher education institutes to act as levers in helping Ireland's economy transition to a knowledge economy. The White Paper "Charting Our Education Future" [13] signposted a role whereby all higher education institutes were given responsibility, through research, for the development of new ideas, new knowledge and the application of existing knowledge. This devolution of accountability and transparency to all higher education institutes has also served to heighten the requirement for timely, accurate and relevant information within Ireland's higher education institutes.

This altered higher education landscape, coupled with Ireland's economic transformation, provided the backdrop to the OECD Review of Higher Education in Ireland in 2004 [14]. Whilst lauding Ireland's success in increasing participation rates, and its growing recognition of the importance of research to sustainable economic growth, the review nonetheless highlighted that Ireland's tertiary education system is at "a significant point of departure" Over the next decade higher education in Ireland will develop against the framework for higher education outlined in the Programme for Government [15] and the National Development Plan [16]. The policy objectives outlined for Ireland's higher education system is to be at the top rank of performance within the OECD [14]. A common feature pervading most of these policy documents is the increasing use of the funding whip-hand to encourage the higher education system to both agree to and implement centrally determined policy initiatives.

More specifically the following development needs are specified as key metrics in the delivery of this policy objective; (1) increased participation and improved access, (2) greater flexibility of course offerings to meet diverse student population needs in a lifelong learning context, (3) enhancement of the quality of teaching and learning, (4) significantly increase the number of PhD numbers and research activity, and (5) effective provision for technology transfer [14]. The Higher Education Authority (HEA) envisions a critical role for the Institute of Technology sector, that is broadly in line with the OECD [14] recommendations, and which recognises: "that the Institutes of Technology and the Universities have developed differently and to establish greater clarity on the diverse role of Institutes of Technology and the Universities in such areas as, teaching and learning, research, regional development, links with industry, social inclusion and meeting skills needs" [14]. The response of the institute of technology sector to these emerging metrics of performance, demanded nationally, yet assessed locally, is the introduction of frameworks of appropriately designed management information systems which seek to;

Improve organisational effectiveness (the determination of unit costs, the quality, retention and destination of graduates, including attrition and retention reporting.).

Understand and minimise the risks facing the respective institute of technology (the generation of strategy supportive data, the prioritisation of data characterised by reliability and timeliness to inform management decision making).

Comply with statutory requirements and corporate governance requirements (the generation of statutorily required financial information and annual reports.).

III. Overview of Information Quality

Currently many of the IQ frameworks are employed post deployment of the IS. A number have been implemented as a direct result of IQ problems, highlighting accuracy or completeness problems in the main. The use of IQ frameworks allows for an examination of the environment without the need for prior knowledge of the development methodology [17].

The area of IQ as a field of research has developed considerably within the past decade, predominantly stemming from reliance by enterprises at all levels on IS for the completion of their core business activities. IS context has also evolved with widespread deployment of web services. Enterprises are now increasingly characterized having mission critical systems based on diverse contexts, yet have no measure of the quality of the information that they derive their business and profit from. This is a wide and varied problem affecting nearly every major type of business activity and enterprise. Al Hakim [6] provides numerous examples of these anomalies from many areas, outlining the reason along with the particular IQ dimension affected. The examples indicate how the generation of information from disparate sources can impact on many aspects of an enterprise, often not even considered when the IS was initially designed. The impact of these IQ problems has prompted researchers to examine such aspects as IQ dimensions, IQ assessment and IQ management.

IV. Design and Develop an IQ Methodology

An examination of the literature [5, 6, 17-22] suggests that the issue of context has not been adequately addressed. Consequently we believe that accessing the same information from diverse contexts will be problematic and inconsistent with existing frameworks and IQ measurement.

Our construction of the IQ methodology examines and builds on work completed by Knight & Burn [20] in the area of algorithm choice for Internet search engines and Naumann and Rolker's [23] analysis of information sources. Our dimension selection were chosen by applying Leungs [22] Importance, Urgency and Cost metric. The methodology proposed is flexible in its application and caters for the employment of other usability metrics such as Doll and Torkzadeh [24] The development of the methodology can be summarised as follows

Identification of the user(s)

Identification of the tasks that make up the system

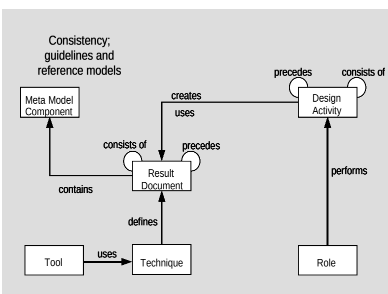
Identification of the dimensions to be assessed

Prioritisation of the dimensions to be assessed

Assessment of the dimensions selected.

In order to overcome these context deficiencies we propose a novel IQ methodology that takes the context of the IS into account, allowing us to analyse the perception of the same information from different contexts. Knight & Burn [20] point out that despite the sizeable body of literature relatively few researchers have tackled the difficult task of quantifying some of the conceptual definitions. Naumann and Rolker [23] further suggest that the subjective view of quality makes the actual assessment of IQ dimensions difficult. The unplanned nature of the evolution with respect to context has the potential to further complicate matters.

The methodology is now illustrated with an example of an IS used by strategic decision makers in the institute of technology sector which is being examined by our research. The primary challenge in the application of the methodology is the analysis of the context and selection of appropriate IQ dimensions.



Role	Design Activity	Result Document	Technique	Tool
Who performs?	What?	With the result?	Applying technique?	Using tool?

Figure 1. Methodology Description

V. Implementation of the Methodology

Grounded IS research to date in the field of IQ has concentrated on intrinsic dimensions. The seminal research with respect to context has focused on user opinion or satisfaction [21, 25].

The most commonly used IS research methodologies have been collated by Galliers [26] and extended by Minger [27]; adopting an interpretive and empirical approach. The primary goal of our research is to examine the relationship between the context of the IS and IQ therefore an empirical approach appears the most appropriate to the research question. This approach allows for an identification of the precise relationship between variables via an experiment design, which can be conducted by laboratory experiments, field experiments or a combination of both.

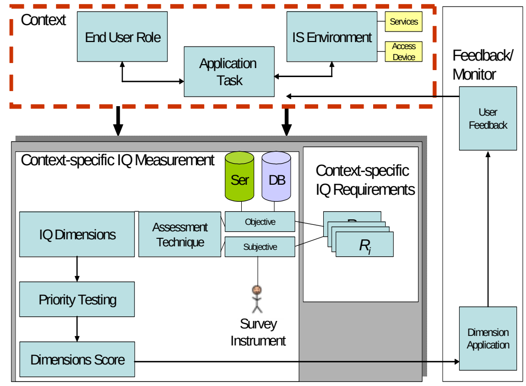


Figure 2. Application of the Methodology

VI. An Experiment

Our initial experiment was conducted over a three month period, involving forty eight participants and three IS at Galway Mayo Institute of Technology. The experiment involved the control of one independent variable namely the access device. Each of the groups were assigned tasks particular to their profile and IS. The tasks within each group were conducted from three different contexts namely workstation, web, and mobile. The first requirement of the methodology is to analyse the software services identified and selected. This at present is a binary test and examines the availability of the service. Prior to the conduct of the full experiment further analysis will be carried out to examine the possibility of applying a more detailed checking algorithm.

The application of the methodology first checked the availability of the three software services identified. In the event of service availability the objective integrity analysers were initiated. The results of this analysis were stored in an IQ table-space similar to an audit table-space [28]. The subjective survey instrument was run in conjunction with the objective integrity analysers. An outline of the user group, task and associated environment was followed by a dimension score for the objective measure and service availability indication. The subjective survey result then followed.

VII. Analysis of Results

The results charts and statistical tests were completed for four dimensions identified by our methodology. We will now describe in detail the process of analysis for one of these dimensions (free of error). The same method and approach was employed in examining all four dimensions.

The initial step in the analysis was the binary test of services. In the pilot study all the services were present therefore the subjective and objective tests were applied to all participants. In order to test the level of significance of the remainder of the results it was necessary to apply an appropriate statistical technique to the data gathered. In the pilot study we sought to ascertain if there was a relationship between the context of the IS (Web, Workstation or Mobile) and the level of IQ. As IQ is a multidimensional concept it was necessary to do this at a dimension level.

The first statistical test completed, illustrated in figure 3 below, is a line plot of the mean between the objective integrity analysis and the subjective survey instrument. Line plots are used to illustrate the frequency of an events occurrence [29]. The pattern on the graph plots the relationship between the objective results for free of error as measured at the database against each of the individual contexts. The graph indicates a relationship between the objective measure and each of the contexts. The objective value is a constant at 89% and the graph illustrates that the subjective measures of the individual contexts follow a distinct pattern. The results clearly indicate that the subjective view of IQ was best from the workstation with less satisfaction from the web and less again from the mobile. In other words the context of the IS access device affects the users subjective view of the IQ. The initial line plot of the mean warrants further examination.

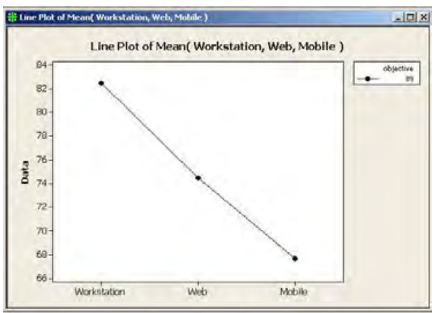


Figure 3. Plot - Objective V Subjective Dimension – Free-of-Error

VIII. ESTABLISH HOW THE PERCEPTION OF IQ IS AFFECTED BY THE CONTEXT OF THE IS.

A variation in the context of the IS we contended impacts on IQ. The context was controlled by manipulating the service level. Similar to the process for objective one we outline this for one dimension in the pilot study and intend to use the same process to complete for all dimensions in the main experiment.

We have chosen the service of systems accessibility. This service is set for various levels and the affect on IQ is measured as per the measurement instruments outlined. Descriptive statistics are applied in the pilot experiment.

This part of the experiment involved varying the levels of systems access and measuring the corresponding IQ dimensions that were identified for the framework.

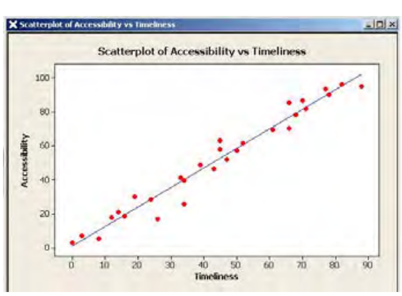


Figure 4. Accessibility V Timeliness

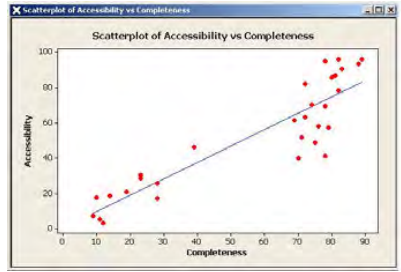


Figure 5. Accessibility V Completeness

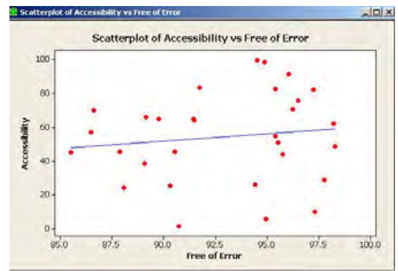


Figure 6. Accessibility V Free of Error

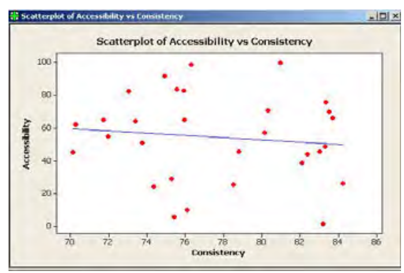


Figure 7. Accessibility V Consistency

The initial results as outlined in figures four to seven indicated that there was a positive linear correlation between system accessibility level and timeliness. As the accessibility levels to the IS increase, the timeliness dimension also improved. Our research identified a relationship between accessibility level and completeness (figure five). At low levels of accessibility completeness is also low, however the relationship does not hold as levels of accessibility improve. At the higher level this relationship was again observed. This fall off in the relationship requires further examination. The scatter plot for both consistency and free-of-error did not display any relationship. We deduced from our experiment that the methodology is sensitive to varying IS context.

Our research also discovered a further significant finding particular to the study of IQ. Many researchers have proposed ways to aggregate single measures of IQ dimensions, often underlying a weighted aggregate of single values for IQ dimensions [25]. Although, recently some researchers have attempted to propose IQ value curves and trade-offs by analysing the potential impacts of IQ [30], many researchers propose to measure the overall impact of IQ as weighted aggregate. A principle measure of the weighed sum of all the criteria (IQC_i) is illustrated below

$$IQ = \sum_{i=1}^{N} \alpha_i IQC_i \quad \text{where} \quad \begin{array}{l} \forall \alpha_i : 0 \le \alpha_i \le 1 \\ \sum_{i=1}^{N} \alpha_i = 1 \end{array}$$

Equation 1 Aggregate measure of information quality

The aggregated value should define the quality level that characterizes information source. The approach to use the average as an aggregation functions may not be suitable among heterogeneous dimensions since dependencies introduces bias that negatively affect the reliability of the assessment procedure. Change in the context (accessibility service) will have an impact on other dimensions in a framework and as a consequence the aggregate score. The research provides a methodology to measure this impact. This it is argued should be completed prior to applying the aggregate, because it will provide a more accurate result with respect to IQ.

IX. LIMITATIONS

Although our research revealed interesting results, this was a pilot study concentrating only on a subset of dimensions and functionality used by strategic decision makers in higher education. However we believe that the principles and methodology proposed have the potential for wider application across IS in higher education. The application of the full experiment will address this limitation. Correlation has been used to examine the relationship between IQ dimensions. Analysis of Variance will be completed for all sections of the research, thus allowing the general application of our findings.

Finally the research to date has only employed experiments along with survey instruments as its research approach. The other data collection methods which will form an integral part of the research such as diaries, semi structured interviews and focus groups will be expanded and form part of the next phase of the project.

X. CONCLUSIONS AND FUTURE WORK

This research contributes to the analysis of IS context and IQ. In recent times companies have invested heavily in IQ programmes in an effort to improve IQ [9]. Our research demonstrates that a relationship does exist between IS context and other dimensions in an IQ framework. The traditional method of measuring IQ dimensions will require re-examination as relationships between the context and other dimensions become more clearly established. By demonstrating clearly that the context of the IS affects the quality of the information produced, strategic decision makers can proactively allow for this when using information generated by these IS. This research also contributes to the field of IQ research by providing a methodology and test environment that can be employed in a context related manner. Upon refinement this methodology has the potential to allow organisations to measure the impact of introducing new contexts post the development of an IS.

REFERENCES

[1] Adelman, S., L. Moss, and M. Abai, *Data Strategy*. 2005: Addison-Wesley.

[2] Date, C., *Databases in Depth*. 2005: O Reilly.

[3] Sangwan, R., et al., *Global Software Development Handbook*. 2007: Auerbach Publications.

[4] Rosenblum, M. and T. Garfinkel, *Virtual machine monitors: current technology and future trends.* IEEE Computer, 2005. **38**(5): p. 39-47.

[5] Lee, Y., et al., *Journey to Data Quality*. 2006: MIT.

[6] AlHakim, L., *Challenges of Managing Information Quality in Service Organizations*. 2007: Idea Group Pub.

[7] Fisher, C., et al., *Introduction to Information Quality*. 2006: MIT.

[8] Gilen, A., B. Waldman, and J. Humphreys, *Technology Assessment: The Impact of Virtualization Software on Operating Environments*. 2006.

[9] AlHakim, L., *Information Quality Management Theory and Applications* 2007.

[10] Mulcahy, *Report of the Steering Committee on Technical Education*. 1967.

[11] *Regional Technical Colleges act*. 1992: Ireland.

[12] *Institutes of Technology Act*. 2006: Ireland.

[13] Education, D.o., *Changing our Education Future*, Education, Editor. 1994, Stationary Office.

[14] OECD, *Review of Higher Education in Ireland*. 2004.

[15] Taoiseach, D.o.A., *Programme for Government 2007-2012*, D.o.A. Taoiseach, Editor. 2007.

[16] Taoiseach, D.o.A., *National Development Plan 2007-2013*, Taoiseach, Editor. 2007, Stationary Office.

[17] Batini, C. and M. Scannapieco, *Data Quality Concepts, Methodologies and Techniques*. 2006: Idea Group Publishing.

[18] Eppler, M. and P. Muenzenmayer. *Measuring Information in the Web Context: A survey of state of the art instruments and an application methodology*. in *5th International Conference on Information Quality*. 2000.

[19] Helfert, M. and M. Ge. *A Review of Information Quality Research*. in *11th International Conference on Information Quality*. 2006.

[20] Knight, S. and J. Burn, *Developing a Framework for Assessing Information Quality on the World Wide Web.* Informing Science, 2005. **8**: p. 159-172.

[21] Lee, Y., et al., *AIMQ: a methodology for information quality assessment.* Information and Management, 2002. **40**: p. 133-146.

[22] Leung, N., *Quality Metrics for Intranet Application.* Information and Management, 2001. **38**: p. 137-152.

[23] Naumann, F. and C. Rolker. *Assessment Methods for Information Quality Criteria*. in *5th International Conference on Infprmation QualityIntetnational Conference*. 2000.

[24] Doll, W. and G. Torkzadeh, *The Measurement of End User Computing Satisfaction.* MIS Quaterly, 1988. **12**(2): p. 259-274.

[25] Wang, R. and D. Strong, *Beyond Accuracy: What Data Quality Means to Data Consumers.* Journal of Management Information Systems, 1996. **12**: p. 5-34.

[26] Galliers, R., *Information Systems Research: Issues, Methods and Practical Guidelines*. 1992: Blackwell Scientific.

[27] Minger, J., *The Paucity of Multimethod Research: A Review of the IS Literature*. 2001, Warwick Business School.

[28] Greenwald, R., R. Stackowiak, and R. Stern, *Oracle Essentials: Oracle Database 11g*. 4th ed. 2004: O Reilly.

[29] Field, A. and G. Hole, *How to Design and Report Experiments*. 2006.

[30] Stvilia, B., et al., *A Framework for Information Quality Assessment. Journal of the American Society for Information Science and Technology*, 2007. **58**(12): p. 1720-1733.

Can Mobile Computing Help Reverse the Decline in Informatics Studies?

D.Vitkute-Adzgauskiene, V.V.Fomin
Dept. of Informatics, Vytautas Magnus University
Vileikos 8,
LT-44404 Kaunas, LITHUANIA

***Abstract*-- In this work we analyse the factors leading to the decline of popularity of traditional Computer Science / Informatics study programmes as observed in Europe and the U.S. We further examine the possible directions for improving the Informatics curricula to reverse the downturn trend. Finally, we report on the first positive experience of including Mobile Computing related topics into the Informatics study programme with resulting increase in students' interest.**

I. Introduction

The last decade saw rapid growth of Internet usage. Our young generation today is usually referred to as the "Internet generation", having a spectrum of uses for the Internet, from a virtual library, to a place for social networking, to study room, to gaming room, etc.

Along with the Internet becoming an inseparable part of our daily activities, phenomena of the "Mobile" came to witness much stronger presence in the life of the world's population. With an estimated 20 percent of the world's population being Internet users in 2007, there were twice as many mobile phone subscribers, and there is a plenty of room for growth, except maybe for the developed countries, where there is already a 100 percent penetration rate [1]. With mobile phones having become a commodity, it shouldn't come as a surprise that mobile broadband is becoming a popular way for accessing Internet, and short messaging and mobile data services are being used to accomplish everyday tasks.

Despite the visible proliferation of ICT (Information and Communication Technologies) in our everyday lives and business, surprisingly, during the last decade universities in Lithuania and globally saw a substantial downturn in Computer Science (CS, referred to as "Informatics" in Europe) and Management Information Systems (MIS) enrollments. In the United States the trend started earlier, with enrollments in CS falling by half from 60.000 in 1998 to 30.000 in 2006/7.[1] In Lithuania, we saw the same 50 percent decreases in admission numbers for Informatics students from 2004 to 2008. While the Lithuanian case can be partly explained by highly hyperbolized admission numbers in 2001-2002 (2-2,5 times higher, compared to 1999) due to the State policy and Government orders, as well as by the possibilities to study abroad after Lithuania joined the EU, there are also "global" factors explaining the decreasing popularity of the Informatics in Europe at large and the U.S. To name a few, the transformation of the hardware (HW) and software (SW) markets, job market trends, transformation of the computing paradigm, the effect of the dot-com bubble burst, inadequate curricula for Informatics education at the school level, etc.

In this work, we reflect on the factors affecting the downturn in Informatics enrollments. We analyze and synthesize data from Lithuania, EU, and the U.S., aiming to understand how the universities can react to the decline in enrollments, and, specifically, whether inclusion of Mobile Computing in the study programmes can help re-instigate students' interest in the Informatics studies.

II. The Rise of the Mobile Society

In many countries, mobile communications have entirely changed the overall national picture of ICT usage. Especially, it is true of countries, which had lower personal computer and fixed Internet penetration before the advance of the mobile era – such is the case for most of the former Soviet Republics, Lithuania being one of these.

Today, Lithuania is a highly saturated mobile phone user country – its mobile penetration by the end of Q2'2008 was 148% [2], i.e., every citizen on average had 1.5 mobile subscriptions in his/her disposition. Besides the population, the number of mobile subscriptions is increasing in machine-to-machine applications (e.g., remote control or security). Demographic categorization reveals that 88 percent of all population, falling into the 16–74 age group, are mobile phone users [3]. For the 16-24 year group this number is even higher – 98%, which let us safely assume that 100% of university students are mobile subscribers.

While fixed Internet user numbers were also growing rather rapidly in Lithuania during the last decade, showing an impressive 835% growth during 2000-2008, the Internet penetration reached only 59 percent of population[2] - much lower than that for mobile.

The relatively low fixed Internet penetration may be the reason why Lithuanian mobile subscribers show high usage of different mobile phone enabled services, as compared to other countries. For SMS, for example, statistical Lithuanian user

[1] CRA Taulbee Survey, http://www.cra.org/statistics/

[2] As of March 31, 2009, according to the Internet World Stats, http://www.internetworldstats.com/stats9.htm.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_25,

sends 156 SMS per month – more than two times the number for the French subscribers, and mobile data services are used by half of the Lithuanian population [2].

Globally as well, mobile Internet is becoming a daily activity for more and more people. In the U.S., for example, the number of people accessing news and information on their mobile devices more than doubled during the last year [4]. Mobile service usage closely follows the habits of modern society. With the growing popularity of social networking services, access to these services from mobile devices has grown over 152 percent in Western Europe from Q3'2007 to Q3'2008. For the U.S., this category saw an impressive 427% increase, while access to news and infotainment services increased by 100-190 percent [5]. For the popular Facebook service, the mobile access rate came to stand even with the fixed Internet access rate.

Interesting to note, the use of mobile services kept increasing even during the economic downturn. The top U.S. sites saw doubled use from wireless pre-paid users over the last year [6], indicating that people consider mobile broadband a convenient and cost saving access channel to everyday (Internet) services.

We can find many examples showing that the mobile usage growth is dependent on the availability of convenient mobile services for everyday use. Mobile parking services in Estonia and Lithuania are a good example – the services gives a possibility of paying for parking, extending parking time, getting corresponding notifications through SMS or mobile Internet services. In Lithuania, mobile payment for parking grew to about 30% of all parking transactions in just 3 years. For Tallinn, the capital city of Estonia, the number is 63%. When looking for what constitutes the "convenience" in these services, in both cases we find that designers of the services were successful in accommodating the habits and preferences of local population (reminders, payment schemes, interfaces, etc.) into the service offering. Other successful non-Internet mobile service uses examples exist, too, such as SMS-based purchase of public transportation tickets, vending machine payments, etc.

The trends, that we notice already, show that mobile phone, possessed today by the absolute majority of active society members, enables fast and convenient access to mobile services. If the necessary services are available, it is rather easy for a user to start using them, and, moreover, to enter the state of "can't do without". However, the growth potential is still high in this area, and challenges exist. With aging population in Europe, and the emergence of "smartphones" as mobile Internet access device, the majority of mobile phone users aren't familiar with many advanced mobile phone features, and would require an adequate training to take advantage of the devices' functionality, as it is getting closer and closer to a computer functionality.

III. The Decline of Informatics at Universities

Despite the growing dependence on ICT in daily lives, work, and communication, for the last decade there was a steady decline in the enrollments to Informatics programmes around the world. When analyzing possible reasons for the decline, we may single out three major factors:

1) Changes in the environment, including technology changes, transformation of the industry, job market's image.
2) Inadequate Informatics education at schools.
3) Conservatism of Informatics study programmes.

A. *Transformation of the industry*

At least since the dot-com bubble burst in 2001, Computer Science (CS) / Informatics enrollments saw a downturn in enrollments, which continues to date. Although slowly recovering since, this time a more fundamental challenge may deepen the "enrollments crisis" – that of the transformation of the managerial and industrial paradigms. This time the new to Lithuania programme of "Business Informatics"[3], and world-established concept and Business Schools' study program of Management Information Systems (MIS) is also under the pressure. The global trend of flattening of enterprise resulted in management functions becoming increasingly horizontal. This, in turn, resulted in elimination of traditional Ford-style vertical integration with its multi-divisional management system. Consequently, where there used to be Chief-X-Officers for every division X, in the new enterprise paradigm the Chief Information Officers (CIOs) found themselves in a situation where they "couldn't communicate with the top management group, were ineffective change agents, or could not contribute to business strategy" [7].

Transformation of the industrial landscape couldn't have left unaffected the job market, and student enrollments in managerial programmes. Secondary school graduates, when deciding on future professions, cannot be encouraged by numbers showing that CS graduates are up to 10 times less likely to be employed, than graduates from any other subject. The employment stats published last year by The Guardian [8] place CS graduates the last in the list of 20 different degree programmes.

The recent data from the U.S. present a somewhat more optimistic picture for the future of Informatics than that in UK. For the first time in six years, the number of CS majors enrolled at U.S. universities increased by 8.1 percent [9]. IT-related jobs are reported to be still in demand and expected to stay at this level or even experience growth in the years to come. Especially this is true for those specialities, that require more complex IT and interdisciplinary skills, e.g. IT management consulting [10].

B. *Inadequate education in schools*

The prospects of the job opportunities as drawn by mass-media in Europe are "complemented" by rather narrow scope of CS/Informatics teaching at secondary schools. This tendency is true both for the U.S. and Europe. For example, a

[3] It is probably correct to say that "Business Informatics" is the closest proxy of Management Information Systems programme in Lithuania.

study by Blum and Cortina [11], as quoted in recent report "Informatics Europe" [12], identified the following reasons for the decline in enrollment in CS courses and programs at the high school level in the US:

- Dot-com bubble burst/not enough high paying careers
- Increased amount of required courses and tests
- Elective status/competition with many other electives
- Lack of understanding of CS by guidance counselors and administrators
- Ill-prepared teachers/lack of interest
- Difficulty of material/not 'fun enough'
- Irrelevance to students
- Lack of meaningful curriculum and standards.

The "Informatics Europe" report [12] further concludes that the situation in Europe is not much better but vastly differs from country to country: "the problem is not only the very limited time for Informatics in the school schedules as compared to e.g. mathematics or physics, but also the lack of good standards for content ('what should we teach'), textbooks ('how do we describe it') and didactics ('how should we teach it') at this level" [12], p.10. Similar to the situation in the U.S., there is a lack of teachers for Informatics in Europe, too.

Besides the narrow scope of Informatics programmes in schools, there is reported to be a pressure on teachers of Informatics from their fellow teachers, nurtured by fear that with the rise of the Internet and ICT, Informatics will present a threat to the established teaching subjects, and also a threat to the teachers of these [12], p.11.

C. *Conservatism of university Informatics programmes*

While the two aforementioned categories of factors contributing to the decline of Informatics programmes are "out of control" for the universities, the revision of study programs can and should be done to tackle the trend.

Looking at the CS/Informatics study programmes at different universities, we see that they hardly make use of the potential of Informatics to cover broad field of study, based on the view of Informatics as the "Science of Information".

The computational aspect in Informatics study programmes is usually addressed to the largest extent, taking into account the large share of the programmes covered by programming, algorithm design, algorithm analysis, data management, operating systems, applied computing disciplines. Cognitive aspects are also to a certain extent present in the majority of the programmes, covered by the mathematical logics, artificial intelligence, neural network disciplines. However, the social aspect issue, becoming more and more important in modern society, is clearly underestimated.

We can observe numerous attempts, trying to overcome the crisis in Informatics studies. A European initiative aiming at taking a timely account of the transformation of the field, is "Informatics Europe" association[4], which was created as a result of the first two European Computer Science Summits (ECSS) held in 2005-6, where heads of computer science departments from all over the European region joined forces to define and promote common policies and study common issues. In August 2007 Informatics Europe published a report "Student Enrollment and Image of the Informatics Discipline" [12]. The report aimed at answering the following questions, among other:

- Why is student enrollment a problem in Informatics/Computing, at least in many of the (Western) European countries?
- Why is enrollment by female students lagging behind? What are the reasons of it, and what can be done about it?
- Do potential students have the right image of Informatics as a field of study, as a science, as a profession?
- How should the field be positioned so it is "clearly" as attractive and challenging as many other disciplines and perhaps even more so?

Their main recommendations on this matter were the following:

- Offer multidisciplinary and cross-disciplinary programs
- Fix computing science's image
- Increase women's enrollment in CS
- Train high school computing science teachers
- Make CS courses fun.

IV. DIRECTIONS OF THE REVISION OF INFORMATICS PROGRAMMES

Some of the recommendations on the revival of Informatics have actually been implemented already for a while in what came to be known as a School of Information (SI / i-School). With some 60 SIs established worldwide up to date, the enrollments of SIs in the U.S. grew 50% since 2001, from 15.000 to over 20.000.

The concept of the School of Information suggests the directions towards which the CS/MIS disciplines should be re-oriented. Specifically, there must be a shift from *isolated* to *integrated*, from *technology-centric* or *community-centric* to *synthesized inter-relational* approaches to informational studies.

Having examined the mission statements of a dozen of the leading North American SIs, the aims and problem areas of the study programmes of different SIs can be synthesized in two broad categories, these of information-centric and technology-centric (see Figure 1 and Figure 2). The former problematization is not that distant from the traditional Informatics studies, and hence the lessons of SIs can be applied to the revival of the discipline.

A. *What Can Be Done At The School Level*

While universities cannot exert direct control over the Informatics' teaching at schools, certain success in stimulating pupils' interest to the studies of Informatics can be achieved with effort from the universities. Making Informatics studies "more fun" starting from the school level could be a way out of crisis. This could be done first of all by

[4] http://www.informatics-europe.org/about.php

adequately exploiting the new trends inside computer science itself: "An unprecedented change in our information processing practices [that] has resulted from the introduction

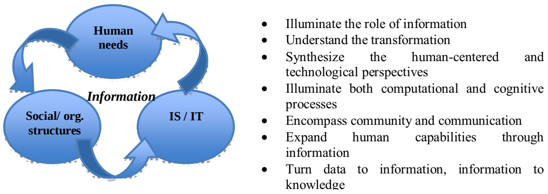


Fig.1. Information-centric approach

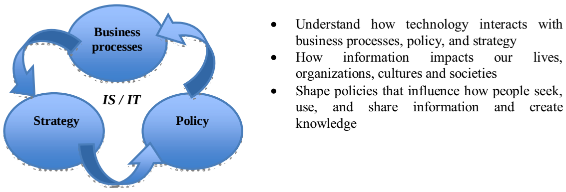


Fig.2. Technology-centric approach

and use of the World Wide Web… [is] being accompanied by technological developments which are in turn educating successive generations as a by product of social practice"[13].

Informatics today is different. It "changes every science that it touches", "it leads the way to many promising new directions in research". And this new image of Informatics must be promoted starting at the highschool / college level, and supported by new educational curricula for universities.

Reports from the U.S. give promising signs on how the task of the raising popularity of Computer Science and Information Systems can be dealt with. One of the reasons for the growing interest in computer science degrees in the U.S. is teens' excitement about social media and mobile technologies. The perception that computer science is cool is spawned by all the interesting things on the Web. The iPhone and Web 2.0 reinforces the excitement, and that attracts the best students [9]. Such excitement could not be not shared by the European teens. However, there may be some directing required on behalf of Informatics' teachers at schools.

B. How Inclusion of Mobile Computing in Study Programmes Can Help Reverse the Decline

Taking into account the transformations in the environment, the rise of the mobile society, and the excitement of teenagers about the social media and mobile, we can expect, that by including the topics of Mobile Computing in the curricula, universities could increase the popularity of Informatics studies. Such an assumption can be to a large extent justified by the success of the i-School concept, teaching programmes of which specifically target the aspect of the interaction of technology with society. Inclusion of Mobile Computing would also reflect the recommendations on the emphasizing the transdisciplinarity in Informatics programmes [13], p.5.

When analyzing the Lithuanian context, we can see, that there are attempts to both add transdiciplinary focus and include Mobile Computing topics in the Informatics study programmes. Most cases observed are related to introduction or development of transdisciplinary Informatics programmes, namely "Business informatics" or "Management Informatics", combining knowledge in the Informatics and Business Management fields. Completely new BSc and MSc programmes in Business Informatics were established at Vytautas Magnus University (VMU) and at Kaunas Humanitarian Faculty of Vilnius University more than a decade ago. Mykolas Riomeris University added a "Business Informatics" BSc programme in 2008, after analyzing its perspectives in the context of other similar programs in Lithuania [14].

The admission numbers for the new business informatics programs show they were popular among students [15], For example, "Business Informatics" MSc programme at Vytautas Magnus University collects twice as many students, compared to the "Applied Informatics" MSc programme. The overall admissions, however, continue to decline.

Besides the introduction of new specializations in the Informatics study programmes, new course modules, related to new business and societal trends can be developed. Here, Mobile Computing comes to play the central role. When talking about mobile communication or ubiquitous computing issues in the university CS curricula, usually two basic opinions prevail. The first of them is stating, that mobile environment is just a subset of a broader electronic environment, therefore this topic shouldn't be considered separately. Another opinion places the mobile computing in a special position, that of boosting the attractiveness and up-to-date characteristics of Informatics study programmes at the universities.

These two approaches to certain extent are reflected in the situation we can see at different universities, offering Informatics study programs. Some have special programmes related to mobile and ubiquitous computing, others have special mobile-related modules included in standard CS programmes, while many remain conservative regarding this point, maybe reflecting only selected topics in corresponding course modules, such as "network management", "Internet technologies", etc.

Looking at the cases when completely new mobile-oriented study programmes are introduced, we mostly find technically-oriented programmes, which are usually categorized as belonging to electronics or electrical engineering fields. Another observed trend here is the orientation towards application design, related to mobile and ubiquitous computing – for example, the MSc Program of the Lancaster university in UK "Mobile and ubiquitous computing" (http://www.comp.lancs.ac.uk/study/pg/ubi/). In this case, two modern computer science trends are combined – the new ways of communicating and accessing the Internet, brought by the mobile computing revolution, and moving away from desktop PCs to everyday digital devices and smart environments. Such an approach can truly make CS teaching programmes more attractive, as it deals with new ways of working, staying in touch, entertaining, and, consequently, with new promising industry and business models.

Talking about mobile-related modules, included into existing study programmes, the majority of them are related either to mobile technology introduction, or to the development of mobile applications. An example of such a course could be “Application development for mobile devices” from Pittsburg University, U.S.A (http://www.ischool.pitt.edu/bsis/course-of-study/course-descriptions.php). We can also notice first attempts include mobile courses, related to the social aspects, as, for example, the “Mobile Social Networking” module, provided by the Department of Continuing Education at the Oxford University in UK (http://www.conted.ox.ac.uk/courses/details.php?id=O07C831H6J).

Looking at the Lithuanian context, we see, that the moderate approach of including separate mobile-related modules or module topics in the existing Informatics programmes prevails - examples include “Mobile Technologies” at Kaunas University of Technology, “Mobile Communication Technologies” at Vilnius Gediminas Technical University, “Mobile and Wireless Communications” and “Mobile Application Development” at Vytautas Magnus University (VMU). Also, mobile-related topics, covering also social aspects, are included in different VMU “Business Informatics” MSc Programme modules, such as “E-commerce and ICT infrastructures”, “Knowledge society ICT strategies”, etc.

The strategy of whether to choose a more revolutionary approach of designing new programmes, or a more moderate approach of improving the existing programmes clearly depends on the country-specific environment, related to the registration and assessment of study programmes, legal base of study regulation, etc. Moderate approach taken by the Lithuanian universities is to a certain extent dependent on the rather complicated current new programme assessment procedure and rather limiting normative documents for Informatics studies.

When evaluating the impact of including Mobile Computing topics in the VMU Informatics study programmes, we can name the following outcomes:

- Elective mobile-related modules attract 2-3 times larger student numbers, than other elective modules offered in parallel.
- The number of students, choosing mobile-related topics for their final thesis, increased 2-3 times (about 20% of BSc Thesis works in 2009 covering mobile aspects).

There are signs of more positive attitude towards the whole MSc programme of “Business informatics” in 2009, the hypothesis based on student responses to different surveys, and quite a number of students expressing their wish to transfer from other programmes. However, the situation will be clarified only after the end of admissions.

V. CONCLUSIONS

In this work we show, that the decline in CS enrollments that is being observed globally can be ascribed to the overly-conservative curricula of the traditional Informatics/CS programmes, which do not reflect the role of mobile computing and social media in the modern society.

The topic of Mobile Computing in CS curricula usually is under-valuated, given the global trend of the growing in scope and depth mobile usage. There are large opportunities of including mobile topics in university CS curricula, such as the increase of attractiveness of study programmes and CS in general, and the possibility to educate specialists with more complex IT and interdisciplinary skills, as well as qualified IT users, thus stimulating the overall knowledge society development process.

The hypothesis of increasing attractiveness of study programmes with mobile-related modules or topics covered, is to a certain extent proven by current examples of Lithuanian Informatics studies, namely by the increased interest in the mobile topics from the electives’ and final thesis perspective, as well as regarding student attitudes towards the whole improved “Business Informatics” MSc programme at VMU.

Two main approaches can be used for including Mobile Computing issues in university programmes – either the introduction of completely new programmes dedicated to mobile topics, or the improvement of existing programmes with mobile-related modules or topics, the choice dependent on the specific country context of study regulation and quality assessment.

REFERENCES

[1] ITU. (2009, 2008-07-15). Market information and statistics (stat): Global ICT developments. 1997-2007. Retrieved Jun 18, 2009, from http://www.itu.int/ITU-D/ict/statistics/ict/index.html

[2] Comm. Regulatory Authority (2009). Lithuanian Telecommunication Sector in 2008 (in Lithuanian). Retrieved June 22, 2009, from http://www.rrt.lt/index.php?-1501980636

[3] Department of Statistics to the Government of the Republic of Lithuania (2008). Information technology usage in households (in Lithuanian). Retrieved June 22, 2009, from http://www.stat.gov.lt/lt/news/view/?id=2737.

[4] comScore Inc. (2009). Mobile internet becoming a daily activity for many. Retrieved June 18, 2009, from http://www.comscore.com/Press_Events/Press_Releases/2009/3/Daily_Mobile_Internet_Usage_Grows

[5] comScore Inc. (2009b). Mobile social networking driving growth of the mobile internet in europe, according to comscore. Retrieved June 18, 2009, from http://www.comscore.com/Press_Events/Press_Releases/2009/1/Mobile_Social_Networking_Europe/(language)/eng-US

[6] comScore Inc. (2009c). Prepaid wireless services attract increased online interest during economic downturn. Retrieved June 18, 2009, from http://www.comscore.com/Press_Events/Press_Releases/2009/6/Prepaid_Wireless_Services_Attract_Increased_Online_Interest_During_Economic_Downturn

[7] Andersen, E. (2007, January 24). Why don't cios become ceos? Retrieved May 11, 2009, from http://www.espen.com/archives/2007/01/why_doesnt_cios.html

[8] The Guardian. (2008, Friday 8 August). Unemployment by degree, 2004-07. Percentage of uk and eu domiciled first degree qualifiers assumed to be unemployed by subject area, 2002-03 to 2006-07. *The Guardian*.

[9] Marsan, C. D. (2009, March 30). Does a computer science degree matter anymore? Retrieved May 11, 2009,from http://www.networkworld.com/news/2009/040609-hot-tech-skills.html

[10] Dawson, C. (2009, March 18). Is computer science the one smart major in a recession? Retrieved May 11, 2009, from http://education.zdnet.com/?p=2271&tag=rbxccnbzd1

[11] Blum, L., & Cortina, T. J. (2007). *Cs4hs: An outreach program for high school cs teachers.* Paper presented at the The 38th SIGCSE Technical Symposium on Computer Science Education.

[12] Leeuwen, J. v., & Tanca, L. (2008). *Student enrollment and image of the informatics discipline* (No. IE-2008-01). Zurich: The Association of Informatics / Computing Departments and IT Research Institutes in Europe.

[13] White, S. A., & Irons, A. D. (2006). The shape of informatics: A UK perspective. In A. McGettrick & G. McAlister (Eds.), *Informatics Education Europe*. Montpellier, France.

[14] Dzemydiene, D. (2008). *The need and perspective for an interdisciplinary informatics study programme* (in Lithuanian). Mykolas Romeris University.

[15] Gruževskis, B., Blažienė, I., Pocius, A., Dikčius, G., Bagdonas, A., Jonauskis, M., et al. (2008). *Systematic assessment of the correspondence between supply and demand of high quality (MSc) specialists* (in Lithuanian). Vilnius.

Towards an Academic Social Network for Bologna Process

João C. Ferreira, Porfírio P. Filipe

{jferreira,pfilipe}@deetc.isel.ipl.pt
GuIAA – DEETC - Instituto Superior de Engenharia de Lisboa, Lisboa, Portugal

***Abstract* — The Bologna Process aimed to build a European Higher Education Area promoting student's mobility. The adoption of Bologna Declaration directives requires a self management distributed approach to deal with student's mobility, allowing frequent updates in institutions rules or legislation. This paper suggests a computational system architecture, which follows a social network design. A set of structured annotations is proposed in order to organize the user's information. For instance, when the user is a student its annotations are organized into an academic record. The academic record data is used to discover interests, namely mobility interests, among students that belongs the academic network. These ideas have been applied into a demonstrator that includes a mobility simulator to compare and show the student's academic evolution.**

I. Introduction: Background and Motivation

With Bologna Process (BP), the social and cultural developments impose a revision of the graduate and postgraduate education systems. Students can personalize their studies from a diversity of options.

Furthermore, study programmes have become modular, giving students the chance of more mobility during their studies. Organizational and legal efforts have been taken to guarantee high compatibility of study programmes and modules, e.g. the so called BP in Europe [1]. Resulting from this process students can personalized their studies and can compose different educational modules at any institutions to a highly personalized curriculum. This process will change the role of educational institutions, such as universities [2]. They are to be considered as components of a common, integrated educational space.

The traditional tasks like offering lectures, seminars, materials, and infrastructure for education will be supplemented by new tasks, such as personalized educational programmes. This scenario requires new information systems and data integration among the different institutions.

In order to support this new academic reality the current work proposes a self-management distributed approach to deal with student's mobility, allowing frequent updates in institutions rules or legislation. The management strategy is based on annotations organized in academic information records, which extend the student profile.

Such system should help current and prospective students to search for suitable course plans or single course units in order to compose a personal curriculum by analyzing their multi-faceted skills. In this scenario, the student proposals for individual studies plan can be submitted and approved by the desired educational institution. The proposed system offers a graphical user interface that is autonomous from a particular institution. The data acquisition and data submission can be implemented in multiple ways such as web services or simple by file transfer protocols.

II. Bologna Declaration

As the Confederation of EU Rectors' and the Association of European Universities (2000) explains it (CRE et al, 2000 p.3-4):

"*The Bologna declaration reflects a search for a common European answer to common European problems. The process originates from the recognition that in spite of their valuable differences, European higher education systems are facing common internal and external challenges related to the growth and diversification of higher education, the employability of graduates, the shortage of skills in key areas, the expansion of private and transnational education, etc. The Declaration recognises the value of coordinated reforms, compatible systems and common action.*

The action program set out in the Declaration is based on a clearly defined common goal, a deadline and a set of specified objectives:

A clearly defined common goal: to create a European space for higher education in order to enhance the employability and mobility of citizens and to increase the international competitiveness of European higher education;

A deadline: the European space for higher education should be completed in 2010;

A set of specified objectives:

- The adoption of a common framework of readable and comparable degrees, "also through the implementation of the Diploma Supplement";

- The introduction of undergraduate and postgraduate levels in all countries, with first degrees no shorter than 3 years and relevant to the labour market;

- ECTS-compatible credit systems also covering lifelong learning activities;

- A European dimension in quality assurance, with comparable criteria and methods;

- The elimination of remaining obstacles to the free mobility of students (as well as trainees and graduates) and

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_26, © Springer Science+Business Media B.V. 2010

teachers (as well as researchers and higher education administrators)."

III. The Social Network

The Social Web represents a set of web sites and applications in which user participation is the primary driver of value. The principal of such systems is well described by Tim O'Reilly [3] and this development is around Web 2.0 [4] principals. Examples of these types of initiatives are Wikipedia, MySpace, YouTube, Flickr, Del.icio.us, Facebook, and Technorati. Discussions of the Social Web often uses the phrase "collective intelligence" or "wisdom of crowds" to refer to the value created by the collective contributions of all these people writing articles for Wikipedia, sharing tagged photos on Flickr, sharing bookmarks on del.icio.us. Knowledge sharing have great potential due to the diversity of information available, diversity of individual cultures and the low costs of gathering and computing over big network (Internet). Currently we have millions of persons offering their knowledge online, which means that the information is stored, searchable, and easily shared, so the big challenge is to find the right match between what is available online and the approach to take the right/useful information from this data. A collective intelligence/knowledge can result from the data collected from all those people that are aggregated and transformed to create new knowledge. This collaborative approach had the spam problem and other fraudulent sources in the mix, but the majority of users have a proper behavior and can handle these disrupt behavior [3].

So, our proposal is collaborative system based on a social network design that is able to connect not only at your own institution but to others within the European Higher Education Space. This collaboration is achieved by users annotations divided into two main description fields: (1) semantic annotation, descriptions using terms from system ontology; (2) descriptions performed by free descriptions terms introduced by the users. All these stakeholders can provide semantic annotation by a specific metadata generation and usage schema aiming to enable new information access methods and to enhance existing ones. A web tool provides access to ontological space, where users can choose predefined terms, see description at section 6. A synergy can be established between students, which are the producers, and the consumers of information and machines in the role of enablers (store, search, combine, inference). In the communication process people create knowledge and machines can help on this process, giving tools to help the process of communication and information sharing. So our approach will address the knowledge offered by millions of people (experts and also novice users) with a background infrastructure for storing, searching and easily sharing of information.

IV. Semantic Web and OWL

The required Interoperability can be achieved by exploiting of Semantic Web Services. Machine readable "semantic" descriptions supplementing Web Services available in a network of cooperating organizations is a remarkable alternative to known standardisation issues. Instead of developing an inflexible standard API and producing a huge overhead concerning its support and evolution, the service provider agree to reference the same domain ontology in their Web Service descriptions. The ability of machine read such descriptions, makes possible the automatic search of required services. Therefore the interoperability of services within the targeted network would be achieved. In this work, one will basically developed reference ontology and the generation of a set of agent applications needed for the search, invocation and orchestration of Web Services offered by educational institutions.

Semantic web architecture is a functional, non-fixed architecture [5]. Barnes-Lee defined three distinct levels that incrementally introduced expressive primitives: metadata layer, schema layer and logical layer [6]. XML and XML schema define syntax, but mean nothing about the data that it describes. That means that some standards must be built on top of XML that will describe semantics of data. This conduct to RDF and a general model in metadata layer RDF schema, which provides mechanisms for defining the relationships between properties declared and others resources. To enable services for the semantic web, one has the logical layer on top. This layer introduces ontology languages, that are based on meta-modeling architecture defined in lower layer. This enables the use of tools with generic reasoning support, independent of the specific problem domain. Examples of these languages are OIL (Ontology Inference Layer), DAML (DARPA Agent Markup Language), OWL and AOBP. OWL is a semantic markup language for publishing and sharing ontologies in the web. OWL is developed as a vocabulary extension of RDF and is derived from DAML and OIL [7, 8].

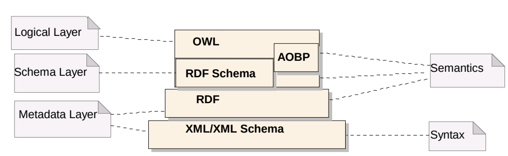


Figure 1: OWL and AOBP in the semantic Web Architecture.

V. Bologna Ontology

The Ontology for Bologna Process consists of four major subjects: (1) Course Plan; (2) Academic Record; (3) Individual Studies Plan; and (4) Diploma Supplement as identified in Figure 2. The Academic Record can be processed to produce a student Individual Studies Plan according to a determined Course Plan. When the student completes the requirements of Individual Studies Plan a Diploma Supplement can be released. Skills as defined in Bologna process are the basis transformation unit.

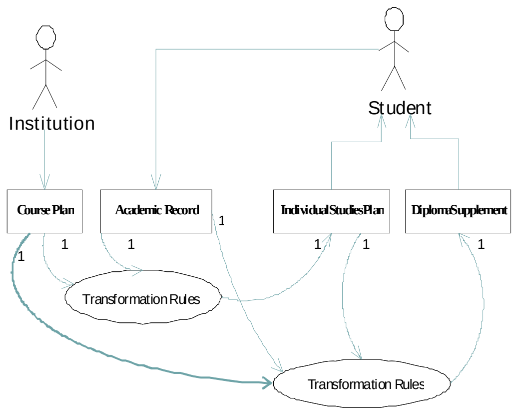


Figure 2: General overview

A. Academic Record

The Academic Record (AR), referenced in Figure 3, can aggregate information items in several languages, namely: (1) Extra Curricular Course; (2) Qualification; (3) A Contract of Studies (is celebrated between the student and the institutions to formalize the terms of the temporary mobility); (4) Work Experience; and (5) Student Data.

The AR includes all student paths during the formation period. There is no need of often cross checks between AR and the information that is kept private by the process stakeholders. For each, information unit updated in the AR, the student should send, if needed, a confirmation document to the institution where he belongs. For instance, if the student completes a professional formation he should update his AR and after he will deliver the right certificate to his higher institution education that holds the AR. When the students do course change or intuition transfer, a copy of the AR is produced and must be detained by the destination institution.

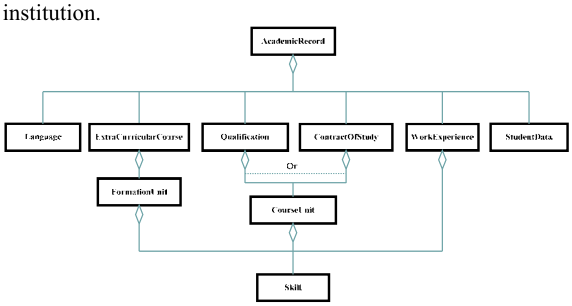


Figure 3: Academic Record diagram

B. Course Plan

The Course Plan characterizes a specific graduation plan by the means of describing the Figure 4: (1) Curricular Structure; and (2) the Studies Plan.

The Curricular Structure represents the distribution of ECTS credits, minimum and maximum, of each scientific area. The goal of the Curricular Structure is to define the amount of ECTS, because of the optional course units. The Studies Plan aggregates all the Course Units that figure in the Course Plan indicating also the skills achieved when the student is approved in the Course Unit.

C. Individual Studies Plan

Individual Studies Plan, illustrated on Figure 5, is similar to a Course Plan representing the already achieved ECTS, according the course units in the student's academic record, and the amount of credits by scientific areas by scientific areas considering a specific course plan.

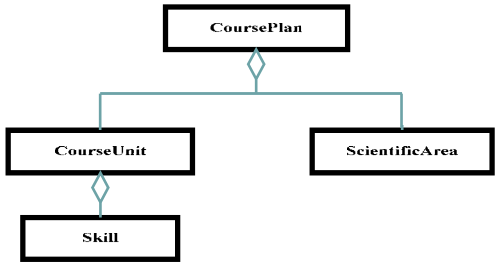


Figure 4: Course Plan diagram

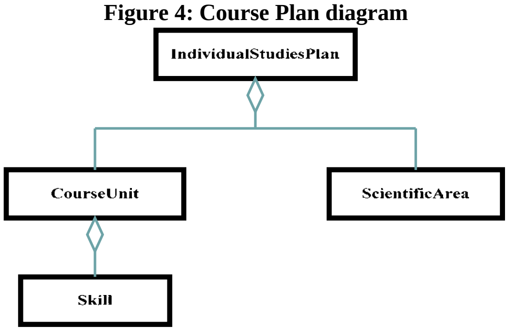


Figure 5: Individual Studies Plan diagram

D. Diploma Supplement

Diploma Supplement follows the model developed by the European Commission, Council of Europe and UNESCO/CEPES. The purpose of the supplement is to provide sufficient independent data to improve the international "transparency" and fair academic and professional recognition of qualifications (diplomas, degrees, certificates etc.). It is designed to provide a description of the nature, level, context, content and status of the studies that were pursued and successfully completed by the individual named on the original qualification to which this supplement is appended. It should be free from any value-judgments, equivalence statements or suggestions about recognition. Information in all eight sections should be provided. Where information is not provided, an explanation should give the reason why. Diploma Suplement diagram is a detailed description of the activities developed by the student in the graduated process based on the follows attributes: employee, student data, course unit, qualification, scientific area and associated language.

This UML representation can transform to a RDF; (2) missing values are replaced with default values when possible, and (3) the RDF is validated against the metadata schema and other validation rules. Each metadata producer gets a report of warnings, errors and other problems that were encountered during harvesting and validating the content. If some parts or all of the metadata is unacceptable due to serious errors, the metadata is discarded until necessary corrections are made. Otherwise, the metadata is added to and published in the Semantic Information Repository (SIR), see section 6.

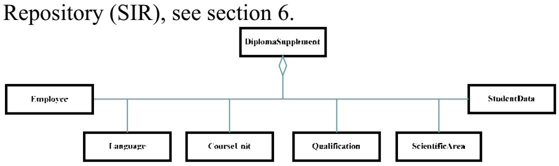


Figure 6: Diploma of Supplement diagram

E. Processing Rules

The skills concept <www.disco-tools.eu> allows the establishment of a semantic link between: (1) Extra Curricular Course; (2) Qualification; (3) A Contract of Studies; and (4) Work Experience.

A specific course plan is completed when the hosting institution recognizes that the student has achieved all the skills that belonging to a course plan.

The skills is a matching unit that allows comparison information items that aggregates its owns skills. A trial example is the comparison of course unit against course unit. However it is possible to compare course units against work experience or qualifications. This matching can be based on historical tables, recommended tables or even requires human supervision.

VI. BOLOGNA APPLICATION SUITE

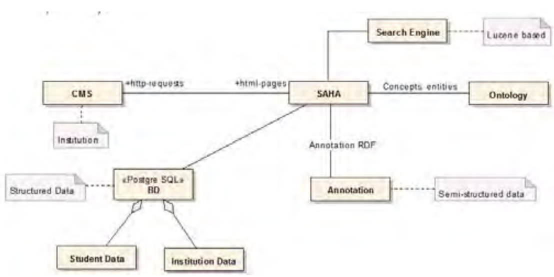


Figure 7: BAS architecture system.

The Bologna Application Suite, is our proposal for a Web 2.0 system environment and major components are described at Figure 7. A major challenge in the distributed content implementation of Bologna process is how to facilitate the cost-effective production of descriptive, semantically correct high quality metadata. In SIR three ways of creating metadata are considered and supported: (1) Enhancing existing web content management systems (CMS) with ontology mash-up service components for producing semantic metadata; (2) Using a browser-based metadata editor for annotating web content; (3) Automatic conversion of metadata. These approaches are outlined below.

A. Enhancing an Existing CMS with Bologna Ontology

Most content providers in SIR use a CMS for authoring, publishing and archiving content on their website. A typical CMS supports creation of textual metadata about documents, such as title and publication time, but not classified annotations. This would require that the system has functionalities supporting ontology-based annotation work, e.g., concept search for finding the relevant concepts (identified with URIs), concept visualization for showing the concept to the user, and storing concepts along other information about the documents. The CMS should also be able to export the metadata, preferably in RDF format, to be used by semantic web applications.

Currently, classification systems are typically shared by downloading them, and each application must separately implement the classified defined schema. To avoid duplicated work and costs, and to ensure that the classified systems are always up-to-date, we argue that one should not only share the classified system, but also the functionalities for using them by centralized mash-up component services. Such services, e.g. Google Maps, have been found very useful and cost-effective in Web 2.0 applications for integrating new functionalities with existing systems.

We have applied the idea of using mash-ups to provide classified system services for the content producers of SIR creating the ONKI Ontology Server framework [9]. ONKI provides ontological functionalities, such as concept searching, browsing, disambiguation, and fetching, as ready-to-use mash-up components that communicate asynchronously by AJAX (or Web Service technologies) with the shared ontology server. The service integration can be easily done by changing only slightly the user-interface component at the client side. For example, in the case of AJAX and HTML pages, only a short snippet of JavaScript code must be added to the web page for using the ONKI services.

The main functionality addressed by the ONKI UI components is concept finding and fetching. For finding a desired annotation concept, ONKI provides text search with semantic auto completion [10]. This means that when the annotator is typing in a string, say in an HTML input field of a CMS system, the system dynamically responds after each input character by showing the matching concepts on the ONKI server. By selecting a concept from the result list, the concept's URI, label or other information is fetched to the client application.

Concept browsing can also be used for concept fetching. Study plan only changed in specific time periods, so the content change can be handling in a better way. In this model, the user pushes a button on the client application that opens a separate ONKI Browser window in which annotation concepts can be searched for and browsed. For each concept entry, the browser shows a Fetch concept button which, when pressed, transfers the current concept information to the client application. ONKI supports multilingual ontologies, has a multilingual user-interface, supports loading multiple ontologies, and can be configured extensively.

ONKI is implemented as a Java Servlet application running on Apache Tomcat. It uses the Jena semantic web framework for handling RDF content, the DirectWeb Remoting (DWR) library for implementing the AJAX functionalities, the Dojo JavaScript toolkit, and the Lucene text search engine.

B. Browser-based Metadata Editor

Some SIR content providers can not add mash-up ontology support to their CMS due to technical or economical reasons. Furthermore, some content providers do not even have a CMS or they may not have access to the CMS that contains the content, e.g., if the content originates from a third party. To support metadata productions in these cases, we have created a centralized browser-based annotation editor SAHA [11] for annotating web pages based on RDF/OWL-based ontology as a schema (collection of classes with a set of properties, described on section 4). An annotation in Saha is an instance of a schema's class that describes some web resource and is being linked to it using the resource's URL (in same cases, URI). We make the distinction between the annotation of a document (e.g. a web page) and the description of some other resource (e.g. a person) that is somehow related to the document being annotated. In addition to containing classes used to annotate documents (annotation classes), an annotation schema used with Saha can also contain reference classes for describing resources other than documents. In other words, an annotation schema forms a basis for the local knowledge base (KB) that contains descriptions of different kinds of resources that may or may not exist on the web. Instances of the reference classes are used as values of properties in annotations. The schema element fields in SAHA can be connected with ONKI ontology service components, providing concept finding and fetching services to the annotator, as discussed above. In practice, an annotation project is Jena's ontology model stored in a database. A model is comprised of the annotation schema and the instances of the schema's classes. It can be serialized to RDF/XML in order to use the annotations in external applications.

C. Automatic Conversion

The third content producing method in SIR is automatic conversion of original data to SIR metadata. This method is currently used in cases where metadata exists in a CMS, but it is in an incompatible format, does not contain ontological annotations (URIs) and/or some minor information is missing in the metadata.

VII. Case Study

We perform a real case implementation of our BAS system to a polytechnic institute of Lisbon (ISEL). ISEL have more than 5000 students distributed by 7 engineering degree (licenciatura in Portuguese) and 6 Masters. Bologna reform was introduced at ISEL in the academic year of 2006/2007. The BAS system was proposed to help academic process regarding students changes (inside institution and outside). System is running for 3 months and we are now getting the first results, with the start of a new academic year. The System prove to use well past experience, i.e. once a course unit equivalence is performed from the course x to the y, the others request are automatic processed. This reality helps a lot because most changes occur from a small percentage of 10 courses. At this moment we have register a community of 100 users (small number) but we think this number will be increase with the usage of the system.

The users when access the system, they need to choose the institution through a menu, see Figure 8. They choose first the country, then the city and finally the institution. Since we did not developed the security module the application is only available in ISEL network, but the BAS system will be prepared for a European usage. Bologna Process makes sense if a big network is constructed around it. Login at BAS can be performed by a Figure 9 menu. If they are new users they can create new information or import academic register from other institution.

Figure 8: BAS System Menu to choose the academic institution.

Figure 9: User login menu: (1) for a new user; (2) to import the Academic Register from other institution.

Due to space limitation is not possible to give all details about the application. For a complete description see [12], we will same detail about collaboration issues, as illustrated on Figure 10 and 11. In the Figure 10, is the menu process associated with the creation of a new qualification. The number two is another collaboration tag where you can make a note to a course simply by being classified and selected the desired course.

Users can create annotations related with transfer process or simply comments about the subject. In Figure 11 we show an annotation process regarding course unit equivalence. Predefine information field's helps the reuse of the information by the users.

A. Simulation of Course transfer

Students simulate a transfer or change of course. A prerequisite to perform this operation is the creation of an individual study plan. To perform the simulation simply select the individual plan of study desired travel destination and the location will be generated an XML document created with all the information of the simulation. The results are represented in a Simulation Menu (Figure 12), where the student can see the results of equivalences if he applies that change.

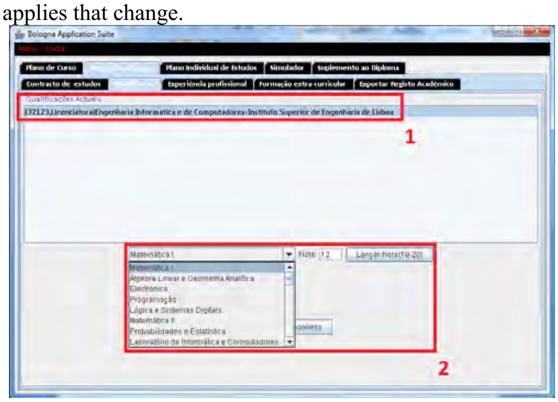


Figure 10: Creating a new qualification.

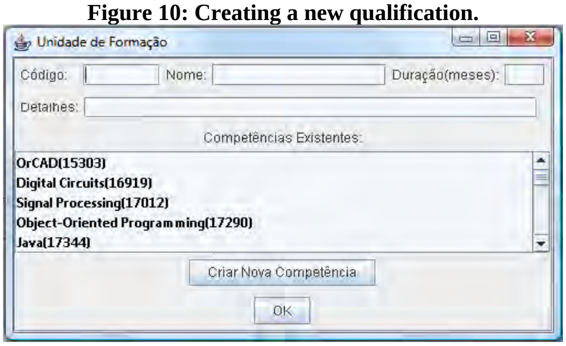


Figure 11: Example of a anotation performed by a student using the descriptors available on the ontological space.

Figure 12: Simulation Menu

VIII. Conclusions

This paper addressed the needs of Bologna Mobility process, with a proposed of BAS system in a Web 2.0 environment to: (1) Support information exchange; (2) Content finding is supported by cross-portal semantic search, based on hierarchical structure imposed by a classified system rather than keywords; (3) The problem of outdated and missing links is eased by providing the end-user with semantic recommendations that change dynamically as content is modified; (4) Content aggregation is facilitated by end-user facets that collect distributed but related information from different primary sources; (5) Quality of content is maintained by including only trustworthy organizations as content producers (academic institutions); (6) share suggestions and comments of students regarding the courses.

At the same time, the problems Bologna Mobility process: (1) Redundancy of content creation can be minimized by the possibility of aggregating cross-portal content; (2) Reusing the global content repository is feasible. By using a central SIR, external applications, can reuse the content; (3) Internal and external link management problems are eased by the dynamic semantic recommendation system of the portal and the content aggregation mechanisms; (4) The tedious content indexing task is supported cost-effectively by shared ontology service components; (5) Metadata quality can be enhanced by providing indexers with ontology services by which appropriate indexing concepts can be found and correctly entered into the system.

References

[1] CRE et al, 2000. Confederation of EU Rectors' Conferences and the Association of European Universities (CRE), 2000. The Bologna Declaration: an explanation.

[2] B.J. Krämer (2003), CUBER für die Ausbildung á la carte, in: G. Simonis (ed.), Lernort Bologna Declaration Communiqué of the meeting of European Ministers in charge of Higher Education,Prague,May 19th 2001, ec.europa.eu/education/policies/ educ/bologna/bologna.pdf.

[3] T. O'Reilly. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, September 30, 2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/ 2005/09/30/what-is-web-20.html>.

[4] Web 2.0 Conferences. http://www.web2summit.com/ Wensing, M., Grol, R. and Smits, A. (1994) Quality judgements by patients on general practice care: a literature analysis. Social Science and Medicine 38, 45±53.

[5] Tim Berners-Lee, Weaving the Web, Orion Business Books, London, 1999.

[6] van Harmelen, F., et al, "OWL Web Ontology Language Reference", W3C Working Draft, http://www.w3.org/TR/2003/WD-owl-ref-20030331/, 2003.

[7] Miller, J., Mukerji, J. (eds.), "MDA Guide Version 1.0", OMG Document:omg/2003-05-01,

[8] Paulo Trigo and Helder Coelho. Agent inferencing meets the semantic web. In Progress in Artificial Intelligence, EPIA-09, volume 5816 of Lecture Notes in Artificial Intelligence (LNAI), pages 497-507. Springer-Verlag, 2009.

[9] Handschuh, S., Staab, S. and Ciravogna, F., S-CREAM -- Semi-automatic CREAtion of Metadata in SAAKM from Web Pages in First International Conference on Knowledge Capture, (2001).

[10] Hyvönen, E., Viljanen, K., Mäkelä, E., et al.: Elements of a national semantic content infrastructure—case Finland on the semantic web. In: Proceedings of the ICSC 2007, IEEE Press (2007).

[11] Valkeapää, O., Alm, O., Hyvönen, E.: Efficient content creation on the semantic web using metadata schemas with domain ontology services (system description). In: Proceedings of the ESWC 2007, Springer–Verlag, Berlin (2007).

[12] Paulos, J. and Costa, R. Bologna Application Suite. Final Year project available at www.deetc.isel.ipl.pt/matematica/jf/bas.pdf.

Supporting Knowledge Discovery in an eLearning Environment Having Social Components

Paola Monachesi and Thomas Markus
Utrecht University
The Netherlands

Vlad Posea and Stefan Trausan-Matu
Politehnica University of Bucharest
Romania

Petya Osenova and Kiril Simov
Bulgarian Academy of Sciences, Sofia,
Bulgaria

Abstract – **One of the goals of the "Language Technology for LifeLong Learning" project is the creation of an appropriate methodology to support both formal and informal learning. Services are being developed that are based on the interaction between a formal representation of (domain) knowledge in the form of an ontology created by experts and a social component which complements it, that is tags and social networks. It is expected that this combination will improve learner interaction, knowledge discovery as well as knowledge co-construction.**

I. INTRODUCTION

In a Lifelong Learning context, learners access and process information in an autonomous way. They might rely on formal learning, that is they might focus on textual material approved by content providers in the context of a course developed by an organization or institution. They might also want to rely on informal learning, that is on (non-)textual content available through the web which is uploaded and accepted by the community of learners and not necessarily by a content provider of an institution.

One of the objectives of the Language Technologies for LifeLong Learning (LTfLL)[1] project is to develop services that facilitate learners and tutors in accessing formal and informal knowledge sources in the context of a learning task. More specifically, the aim is to support learner's interaction in order to facilitate knowledge discovery and knowledge co-creation. To this end, a Common Semantic Framework (CSF) is being developed. The CSF allows the stakeholders to identify, retrieve and exchange the learning material. More specifically, it supports formal learning by addressing course material which includes textbooks, articles, slides as well as informal learning which we identify with (non-)textual material emerging from social media applications. Within the CSF, communication is facilitated through the use of social networks and new communities of learners are established through the recommendations provided by the system. In order to provide recommendations, the user's profile, his interests, his preferences, his network and obviously the learning task are taken into account. It is through the access to formal and informal material that new knowledge will originate.

Knowledge repositories are employed to achieve the above goal. An appropriate way to retrieve formal content might be by means of an ontology which can support the learner in the learning path, facilitate (multilingual) retrieval and reuse of content as well as mediate access to various sources of knowledge, as concluded in [1]. In our specific case we use an ontology in the domain of computing, which is employed to provide deep annotation of the learning materials that should facilitate their understanding and reuse.

However, the ultimate goal of the LTfLL project is to complement the formal knowledge represented by ontologies developed by domain experts with the informal knowledge emerging from social tagging in order to provide more flexible and personalized ontologies that will include also the knowledge of communities of users. The enhanced ontology will be connected to the social network of the learners through the tags that they provide, improving thus the possibility of retrieving appropriate material and allowing learners to connect to other people who can have the function of learning mates and/or tutors. In this way, it will be possible to provide a more personalized learning experience able to fulfill the needs of different types of learners.

The paper is organized as follows. In section II, we discuss the differences between formal and informal learning. Section III presents the design of the Common Semantic Framework, while section IV focuses on its implementation. In section V, we present an initial evaluation of the system while VI introduces some concluding remarks.

II. FORMAL VS. INFORMAL LEARNING

As already pointed out above, in the LTfLL project, we are developing services that facilitate learners and tutors in accessing knowledge sources within the context of a certain learning task. The learning process is considered an attempt to make the implicit knowledge which is encoded in the learning materials, explicit.

In the LTfLL project, we aim at supporting both types of learning - formal and informal one. Under 'formal learning' we understand the process, in which the learners follow a specific curriculum. For each topic in the curriculum, there is a set of learning materials, provided by the tutor. In contrast to this, the curriculum in the informal learning process is not obligatory, and the role of the tutor either is absent or is not obvious. The learners exploit different sources to succeed in their learning goal, very often relying on the information, available on the web, and on the social knowledge sharing networks.

From the perspective of formal learning (i.e. guided, monitored one), learners rely mainly on material which has been prepared by the tutor on the specific topic and is addressed by the curriculum. Learners are expected to access certain pieces of pre-stored information, which would lead

[1] http://www.ltfll-project.org/

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_27, © Springer Science+Business Media B.V. 2010

them to the required knowledge. In the informal learning process, learners have to locate knowledge sources on the web, then to select the relevant ones, and finally to investigate each of them. During this process, learners often need guidance and thus they might profit from finding appropriate people (peers) on the web.

The two learning paradigms are complementary from the lifelong learner's perspective. In the LTfLL project, we are developing two sets of services that would support both - the formal style of learning and the informal one. These services are integrated within the Common Semantic Framework.

III. COMMON SEMANTIC FRAMEWORK: DESIGN

One of the aims of the LTfLL project is to build an infrastructure for knowledge sharing which we call Common Semantic Framework (CSF). It allows for identification, retrieval, exchange and recommendation of relevant learning objects (LOs) and of peers. It is ontology driven allowing thus for a formalization of the knowledge arising from the various stages of the learning life-cycle. Domain ontologies offer useful support in a learning path since they provide a formalization of the knowledge of a domain approved by an expert. In addition, ontologies can support formal learning through the annotation of the learning objects [1].

More specifically, with respect to formal learning, we are developing a Formal Learning Support System (FLSS) that supports the individual work of the learners and tutors in manipulating the learning objects, including the knowledge which is implicitly encoded in them and in adding their own descriptions to these learning materials. In particular, an annotation service has been implemented which is meant as a way to make the implicit knowledge attested in learning objects, explicit. Thus, through the annotation service, we aim at relating the knowledge, encoded in the text, to the formally represented knowledge in a domain ontology.

Despite the possible limitations of the ontologies, they provide a highly structured way of interlinking different knowledge sources. From the point of view of the learning process itself, the domain ontology can be viewed as a way to produce a concise representation of the most important concepts in a given domain. Consequently, the learner has to interact with them in order to assess his current level of familiarity with the domain or to use them to acquire new knowledge.

While in formal learning, courses and learning objects play a relevant role, in informal learning, communities arising from social media applications are becoming more and more dominant, but are poorly integrated in current learning practices. Learners prefer to use the internet to search for an answer rather than asking a tutor or a colleague [3]. Most of the documentation for a given learning task is found on the internet and most learners do not check if the information is accurate or reliable [4]. Recently, the usage of the internet has shifted towards social networking applications. 1 billion movies are seen everyday on YouTube.com[2] and 150 million users are logging each day to Facebook.com.[3]

The communities that emerge through social media applications such as Delicious, Flickr or YouTube provide two important elements that can be integrated in the Common Semantic Framework, in order to support informal learning. They provide the "knowledge of the masses" in the form of tags that are being used to annotate resources. On the other hand, the structure of these communities which can be represented through a social network can be employed to recommend content and peers. The social networks automatically introduce a notion of trust in the search results in relation to the user's social network.

The knowledge produced by communities in the form of tags can be employed to enrich existing domain ontologies semi-automatically. More specifically, we have merged the dynamic knowledge provided by users/learners through tagging with the formal knowledge provided by the ontologies by adding tags/concepts (or instances of concepts) and relationships between concepts in the domain ontology. In the CSF, the connection between words/tags and concepts is established by means of language-specific lexicons, where each lexicon specifies one or more lexicalizations in one language for each concept. DBpedia [5] is used as a knowledge base to resolve the extracted lexicalization to unambiguous existing concepts. DBpedia doesn't provide much information about relations among concepts when compared to specially crafted domain ontologies, but compensates this shortcoming with the huge number of available concepts.

There is an important side effect of this ontology enrichment process: if tags given by learners or emerging from social media applications are related to the concepts present in the ontology, we manage to include not only the expert view of the domain, but also the learner's perspective. In fact, domain ontologies developed by knowledge engineers might be too static, incomplete or might provide a formalization that does not correspond to the representation of the domain knowledge available to the learner which might be more easily expressed by the tagging emerging from communities of peers via available social media applications.

It is important to help the learner manage the content that is created daily around him in his learning communities. The knowledge produced inside the learner's network is very valuable to him. According to the Communities of Practice (CoP) theory [6], a learner acquires knowledge, as he moves from the periphery to the center of a CoP.

In order to facilitate the informal learning experience, we automatically monitor the learners' activities and his peers on the social network applications. The data that the learner and his network create or annotate is indexed in a semantic repository. This data can then be used to offer learning support which exceeds that which is currently offered through keyword search or simple recommendations.

In this way, the CSF supports self-organization and the emergence of collaborative knowledge and classification. In addition, we aim at connecting learners to other learners in an

[2] http://youtube global.blogspot.com/2009/10/y000000000utube.html

[3] http://www.facebook.com/press/info.php?statistics

appropriate way. To this end, the content the learner is searching and selecting can be used as a trigger to get him in touch with other users that employ similar content or annotations. Furthermore, the system monitors the changes that appear in his network with respect to content and to users over time and will provide recommendations for how he could update his network (by adding new peers or removing old ones). Managing the social network is especially important for novices that need to create their own CoP with people with a similar interest. Alternatively, if a learner is already part of a CoP, he needs to keep up with the changes or new information in his domain(s) of interest. Most of the learner's attention will probably go to content produced by peers that are highly relevant and trustworthy. Therefore, it is important that the system maximally exploits these relations.

To summarize: In the CSF we establish an obvious link between the network of users, tagging and the resources (cf. also [7]). We further extend this theoretical model by automatically linking and enriching existing domain ontologies in order to structure the heterogeneous information present in social networks, (deeply) annotated resources, and concepts. The knowledge rich ontologies integrated with tagging provide compelling advantages for a wide range of tasks such as recommendation, community building and discovery learning.

IV. COMMON SEMANTIC FRAMEWORK: IMPLEMENTATION

The Common Semantic Framework implementation is based on a Service Oriented Architecture (SOA) that represents the backbone of the system. It includes five components: data aggregation, ontology enrichment, multimedia document annotation, search and visualization. The services that link these components have a strong emphasis on semantics by employing shared URI's for all the various entities such as concepts, resources, users and annotations. Furthermore the information exchanged through the Web services carries semantic meta-data, as most of the data used in the CSF is stored in RDF and OWL.

As can be seen in figure 1, the communication between the modules is implemented using web services and the core of the CSF is represented by the semantic repository. The repository adopts RDF/OWL semantic vocabularies like SIOC, FOAF, Dublin Core and SKOS to represent information extracted from the social networking applications and employed by all the modules. The other modules, that is ontology enrichment and search, use web services to extract information from the repository in order to provide the required information. The final results are converted to data that can be visualized through a widget inside a learning or social networking platform (e.g. Moodle or Elgg).

Here, the various components are described in more detail.

Data Aggregation

The data aggregation module consists of a crawler and the services that link it to the semantic repository. The crawler uses APIs provided by an increasing number of social networking applications to get information about users, resources and tags.

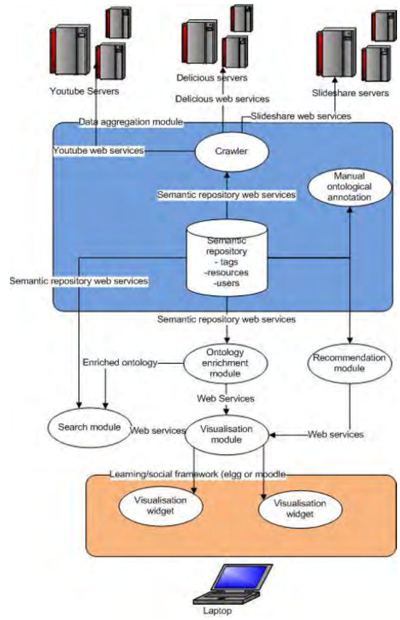


Figure 1. The Common Semantic Framework implementation

The crawler extracts information from a user's social network such as indexing bookmarks that the user is posting on Delicious, videos posted on YouTube, slides posted on Slideshare together with the tags used to classify the resources and information about the social connections developed inside these web sites. The data is converted into semantic information using ontologies like SIOC, SCOT, FOAF.

The data extracted by the crawler can be interpreted as a folksonomy, which is a hypergraph describing the information about users, resources and tags as specified in [7]. The storage of the folksonomy in a repository allows the system to provide various search and recommendation services. These services can provide references to users, resources, tags, concepts and annotations that the learner can use in his learning activities. The advantage of using a semantic repository and commonly used vocabularies for the purpose of data integration is that we obtain a highly integrated dataset containing semantically compatible data from different internet sources. These data can be used to enrich an existing domain ontology for providing improved search services that can exploit the structured knowledge formalized within an ontology as well as the knowledge coming from the participating communities.

It should be noted that the repository also contains a collection of learning material collected from the web which has been semantically annotated.

Ontology enrichment with social tags

The purpose of this module is to benefit from the information extracted from social media applications for the enrichment of existing domain ontologies. To this end, we determine which concepts and relations are the best candidates to be added to the given domain ontologies. In the CSF we take, as starting point, the ontology on computing that was developed in the "Language Technology for eLearning" project (cf. [1]). It contains 1002 domain concepts, 169 concepts from OntoWordNet and 105 concepts from DOLCE Ultralite. Candidates are generated by applying similarity measures or other sorting mechanisms to derive socially relevant tags. Implemented examples include coocurrence-based and cosine-based measures [9]. This analysis drives lookups and selective integration from DBpedia for the domain ontology enrichment. The various web services provide additional lexicon entries, socially relevant specific classes linked to more abstract classes already present in the ontology and additional relationships discovered between already existing domain concepts. The system provides an automatically enriched ontology which contains the vocabulary of the Community of Practice that the user is part of. More specifically, the resulting ontology integrates the socially relevant concepts, but with the structure of an expert view domain ontology. Extraction methods exclusively focused on deriving ontology-like structures from folksonomies cannot provide such a high quality of results due to unavailability of the implicit knowledge in folksonomies which has been made explicit in domain ontologies.

Multimedia Document Annotation

The aim of the multimedia document annotation is to annotate explicitly learning material with concepts from the domain ontology previously mentioned. The purpose of this annotation is to allow semantic search, based on the ontology. The usage of the domain ontology in the search provides abstraction from the concrete wording of. The annotation is done in the document. This type of annotation allows the user to be directed to a particular part of the document.

Within the CSF both text annotation and image annotation are possible. The text annotation is implemented as a language pipe performing tokenization, POS tagging, lemmatization, semantic annotation and coreferential annotation. The first three steps rely on available state-of-the-art tools. The semantic annotation is performed by a concept annotation grammar constructed on the basis of an ontology-to-text relation as described in [13] and [14]. The coreference annotation module provides mechanisms to make the concept annotation more precise and with a better coverage of text (15% improvement).

The image annotation is performed manually. It is complementary to the text annotation. The reason to include it is that very often the pictures and diagrams, etc, present in the learning documents contain important information for the learner. The user interface consists of a document browser, image editor and ontology browser. The user selects an image from the document browser. The image is opened within the image editor. Then the user can select a region of the image and assign one or more concepts to the region. The interpretation of this type of annotation is that the region depicts an instance of the corresponding concept.

The document annotation is stored in the semantic repository together with links to the documents which will be used for document search by the ontology search engine.

Search

The search service takes the learners' query as input and analyzes the information stored in the semantic repository in order to provide relevant and trusted results. At the moment, we have implemented two types of search, that is ontology based search that exploits the structured information attested in the ontology as well as search based on tags and social networks that exploits the social aspects of the knowledge provided. Our plan is to integrate these two types of search into a recommendation system. It should exploit the strong features of both types of search, that is ontological structure and community based knowledge (i.e tags and social networks).

Ontology based search

The semantic search can be triggered by a list of keywords, provided by the user. The query string is analyzed on the basis of a lexicon, mapped to the domain ontology which is enriched with tags. The result of this analysis is a set of concepts from the ontology. The concepts, related to the keywords, are additionally processed on the basis of the information from the ontology by adding related concepts, that is query expansion. Query expansion is done via reasoning over the ontology. The most typical case is the addition of the sub-concepts to the ones extracted from the ontology on the basis of the keywords. The initial set of concepts is interpreted conjunctively and the addition from the query expansion is added disjunctively to it.

The query, formed in this way, is evaluated over the repository of annotated documents. The result is a list of documents and in the case of the annotated ones it is a set of pointers to parts of the document which are considered as relevant to the query - sentences, paragraphs and pictures. The possibility of ontology browsing is also offered to the learner. More specifically, the enriched ontology can be exploited by the learner to get an overview of the domain specifically tailored to some search query, relevant user or specific document. This allows the learner to discover resources which do not match previously known keywords by supporting discovery of new concepts and associated resources through a visualization of the domain.

Search based on social networks and tags

In addition to ontology based search and browsing we have implemented an alternative type of search based on the structure of the social network and the tags provided to resources.

More specifically, in order to return trusted results, the search service focuses on documents created or bookmarked by the learner's friends and close contacts. The results found in this way also contain the name of the users who created or recommended them. Being in the learner's network, he can now contact them and ask for more information or simply trust the document returned more as it has the guarantee of a peer recommendation. The resources found this way will most likely be in the learner's zone of proximal development [10].

The search algorithms can also identify the most relevant peers for a given task. The learner can interact using chat or e-mail with the peer to ask further questions and the peer might very well be helpful. He is already part of the learner's social network and we suppose that this means there is already a connection established between them.

The search services use the FolkRank algorithm [8]. This algorithm is used because it can retrieve any type of content that has been previously tagged.

A similar tag-based algorithm is used to search documents on the basis of the tags provided by users [11]. The algorithm computes results based on tag co-occurences, on tag-resource and on tag-user affinities. These affinities are computed based on the frequency of the user-resource and tag-user pairs. The tag co-occurence is the most used measure for tag similarity and it measures how often 2 tags are used together. After computing the co-occurence matrix and the affinities matrices, the algorithm clusters tags and computes an affinity between a user and a cluster of tags. Finally, using the affinity between user and the tag clusters the algorithm returns a number of resources of interest.

Visualization

The visualization module offers a graph based view of either a social network of users and resources centered around a given learner or a part of a domain ontology.

The social network view can present the whole network as it exists in the repository or it can show only the part of the network returned by a search query. The search results are presented as a network of users and resources attached to them. The user can thus see which path links him to a specific user or resource on the given search topic. In this way, the user can identify the best means to contact a user for further questions and can also rapidly identify who are the most competent users on a specific topic. The unified relations represented graphically are from peers across different social networking applications because our system monitors and integrates all the social media applications that the learner is using.

The ontology based graph allows a user to quickly gain insight into the important relations which exist between domain concepts. The ontology visualization is a hyperbolic hypergraph that contains domain concepts with their preferred lexicalisation and (taxonomic) relations to other domain concepts that are drawn as labeled edges. The hyperbolic geometry allows the user to focus on the important topics while still retaining a sense of the larger surrounding context. The user can discover, previously unknown, concepts which would not have shown up either in the social network or through a keyword search, by interactively browsing the ontology.

Together the visualization services provide a way for the user to navigate through the knowledge sources, focus on specific details, add new information, take notes, etc.

The visualization can be customized to be included in a number of widgets to show specific information like visualizing a part of the ontology, the social network, a definition, relevant peers, etc. These widgets can easily be recombined or disabled in order to provide a learning environment that best fits the user.

V. Evaluation

The Common Semantic Framework was evaluated on the basis of three use cases:

1. tutor support in the creation of a new course;
2. comparison between browsing of an ontology enriched with social tags and a cluster of related tags in order to solve a quiz;
3. recommendation of content and peers on the basis of a social network and tags.

In the first use case, the scenario employed for the validation focused on providing support to tutors in the creation of courses in the IT domain. The idea was to evaluate the efficiency of the computer-assisting tool when creating a course on a certain topic. This evaluation focused on the performance of the various functionalities, reflected in users' opinions. The outcome would give feedback to the developers on how to improve the supporting system. The evaluated functionalities were the following: relevance of the learning material corpus wrt a domain topic; relevance of the retrieved material; suitability of the ontology for structuring a lecture in the chosen domain; possible visualization of the interface; combination between ontology browsing and semantic search.

There were two target groups: five teachers and three managers (vice-deans at the Faculty of Slavonic Languages, Sofia University). Our assumptions were as follows: the combination among semantic search, ontology browsing and a concept map based visualization would support the tutors in the processes of finding the relevant materials, structuring the course and representing the content in a better way. Our hypothesis was that, on the one hand, the tutors might differ in their preferences of using the functionalities. For some tutors, search would be more intuitive, while others might prefer structure (in our case - ontology). On the other hand, the managers would look at the facilities from a different perspective, such as how this supporting system would communicate with the well-known Moodle, for example. Two methodological formats were adopted: the think-aloud strategy for the tutors, and the interview for the managers. The think-aloud format included seven questions, related to the specific functionalities (search, browsing, visualization) as well as the general usability and problematic issues of the system. The interview format included three questions, which were related to the usability and management issues related to the adoption of the system.

The results confirmed our hypotheses. Additionally, four teachers reflected in their answers the fact that the functionalities are more efficient when operating in some interconnected mode. Regarding the semantic vs. text search three tutors gave preference to the semantic one, while two were more cautious and balanced between the two. Visualization was the most commented functionality, because the teachers had different ideas on how it should be controlled. The main worry of the managers was the requirement of a training course to be able to use the CSF.

In the second use case, we have focused on the support provided by an ontology enhanced with tags in comparison with clusters of related tags extracted from Delicious (lacking ontological structure) in the context of a learning task. The underlying assumption being that conceptualization can guide learners in finding the relevant information to carry out a learning task (a quiz in our case). The hypothesis was that learners might differ with respect to the way they look for information depending on whether they are beginners or more advanced learners. While beginners might profit from the informal way in which knowledge is expressed through tagging more advanced learners might profit from the way knowledge is structured in an ontology. The experiment was carried out with six beginners and six advanced learners, all with an academic background. Responses were elicited by means of a questionnaire that contained questions about which elements learners used to provide an answer to the quiz (i.e concepts, relations, documents, related tags, structure).

The responses to the questionnaire by beginners showed that the enriched ontology can be a valid support to answer certain types of questions, because the relations between the concepts are given in an explicit way. However, beginners rely mainly on documents to find the relevant information both in the case of the enhanced ontology and in the case of the cluster of related tags. This attitude is probably influenced by the way search occurs in standard search engines.

On the other hand, advanced learners indicated that they were able to find the answer to the quiz quickly by using the enriched ontology and they also used less documents to find the answers. They gave the structure of the ontology a higher rating than the beginners. Interestingly, advanced learners were more positive about the structure of the tag visualization than the beginners. This is probably due to their background knowledge that enabled them to interpret the graph better. We refer to [12] for additional details on the results of the experiment and their interpretation.

In the third use case, we have focused on how we can support the activity of a group of learners (six) and one tutor that are using social networking applications for learning purposes. The tutor looked for a way to recommend resources for the learners in a time efficient way while the learners needed trust-worthy documentation on a given topic. The idea of the experiment was to see if the combined usage of social networking environments together with our tag based search could be valuable for their learning experience.

The scenario was the following: the tutor and the learners provided only their public usernames of their social networking sites to our system and continued using the platforms as before. The crawling service gathered information about them, their network of friends and the resources present in these networks. The learners searched for documents of interest to them and the system returned documents written or recommended by their peers ordered by the relevance of the content. Among the results were some documents bookmarked or written by the tutor. The results contained the name of the document together with the name of the peer that recommended the document. The learners trusted the resources retrieved this way more than the ones retrieved on a similar Google search due to the fact that they knew and trusted the people behind those resources. For the tutor it was even easier, as his task consisted only of creating and bookmarking resources. As the learners were already connected with the tutor, his resources were part of their search results. That means they could get answers from their tutor without actually bothering him.

The only inconvenience with this scenario for our little experiment group was that the tutor was not comfortable with all of his resources being accessible to the learners. These privacy issues have to be taken care of in the next version of the platform.

VI. Conclusions

One of the goals of the LTfLL project is to develop services that facilitate learners and tutors in accessing formal and informal knowledge sources in the context of a learning task. To this end, a Common Semantic Framework is being developed that provides recommendations on the basis of the user profile, his interests, his preferences, his network and the learning task. An advantage of this system with respect to standard search (e.g. Google) is that the strong semantic component behind the search and recommendation system as well as the integration of social network analysis will improve learner interaction, knowledge discovery as well as knowledge co-construction.

References

[1] P. Monachesi, L. Lemnitzer, K. Simov, "Language Technology for eLearning", in *Proceedings of EC-TEL 2006, in Innovative Approaches for Learning and Knowledge Sharing, LNCS 0302-9743,* 2006, pp. 667-672

[2] L. Lemnitzer, K. Simov, P. Osenova, E. Mossel and P. Monachesi, "Using a domain ontology and semantic search in an eLearning environment", in *Proceedings of The Third International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering. (CISSE 2007).* Springer-Verlag. Berlin Heidelberg, 2007

[3] V. Posea, S. Trausan-Matu, and V. Cristea, "Online evaluation of collaborative learning platforms", in *Proceedings of the 1st International Workshop on Collaborative Open Environments for Project-Centered Learning, COOPER-2007, CEUR Workshop Proceedings 309*, 2008

[4] L. Graham , P. Takis Metaxas, ""Of course it's true; I saw it on the Internet!": In critical thinking in the Internet era", *Communications of the ACM*, vol. 46, May 2003, p.70-75

[5] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak and Z. Ives, "DBpedia: A nucleus for a web of open data", in *Lecture Notes in Computer Science 4825*, Aberer et al., Eds. Springer, 2007

[6] E. Wenger, *Communities of Practice: Learning, Meaning, and Identity*, Cambridge: Cambridge University Press, 1998

[7] P. Mika, "Ontologies are us: A unified model of social networks and semantics". *Journal of Web Semantics* 5 (1), page 5-15, 2007

[8] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "FolkRank: A Ranking Algorithm for Folksonomies", in *Proceedings FGIR 2006*, 2006

[9] C. Cattuto, D. Benz, A. Hotho, and G. Stumme, "Semantic analysis of tag similarity measures in collaborative tagging systems," May 2008. [Online]. Available: http://arxiv.org/abs/0805.2045

[10] L.S. Vygotsky, *Mind and society: The development of higher psychological processes*, Cambridge, MA: Harvard University Press, 1978

[11] S. Niwa, T. Doi, S. Honiden, "Web Page Recommender System based on Folksonomy Mining for ITNG '06 Submissions", In *Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)*, 2006, pp.388-393

[12] P. Monachesi, T. Markus, E. Mossel, "Ontology Enrichment with Social Tags for eLearning", in *Proceedings of EC-TEL 2009, Learning in the Synergies of Multiple Disciplines*. LNCS 1611-3349, 2009, pp. 385-390

[13] P. Osenova, K. Simov, E. Mossel "Language Resources for Semantic Document Annotation and Crosslingual Retrieval". In: *Proc. of LREC 2008*, ELRA.

[14] K. Simov and P. Osenova "Language Resources and Tools for Ontology-Based Semantic Annotation". *OntoLex 2008 Workshop at LREC 2008*, pp. 9-13

Computer Numerical Control, Mass Customization and Architectural Education in an Emergent Economy

Neander F. Silva[1]
Faculdade de Arquitetura e Urbanismo, Universidade de Brasília, Brasil

Alan H. Bridges[2]
Department of Architecture, Strathclyde University, United Kingdom

Ecilamar M. Lima[3]
Faculdade de Arquitetura e Urbanismo, Universidade de Brasília, Brasil

***Abstract* – The objective of this paper is to demonstrate that digital fabrication has had little impact in the Brazilian architecture so far, as compared to other countries, not because of an alleged unavailability of CNC technology in this country's construction industry, but for other reasons that must be investigated. We show in this paper the results of a preliminary survey carried out in the region of Brasilia that reveals a significant presence of CNC technology in the local construction industry which points to new opportunities for innovation in the architectural field. We will also show that some experiments in architectural education involving rapid prototyping and digital fabrication may provide a lead towards the solution of this problem.**

I. Introduction

Computer resources that allow the computerized manufacturing of artifacts from three-dimensional virtual models have been in use routinely by the car and aircraft industries for several years. We will refer to such resources from now on as digital fabrication.

These digital resources, if incorporated to the process of the built environment production, allow a fundamental paradigm change to take place in contemporary architecture. The construction industry has relied so far in the mass production of standardized components. The produced elements are generic material which will be customized later in the life cycle of the product. The mass produced components are classified into defined categories and produced in a limited array of forms and sizes. They are then stored and catalogued until they eventually, if sold, end up in a combination of elements in a factory or as a part of a building in the construction site [1].

As a new paradigm, the mass customization provided by such digital resources, allows that construction components may be produced for specific purposes, to become singular elements in unique contexts of specific buildings. The savings obtained in the automation of this process mean that the costs of singular components are hardly different from the old standardized ones [2][3][4][5].

(1) neander@unb.br, neander.furtado@gmail.com, http://lecomp.fau.unb.br
(2) a.h.bridges@strath.ac.uk, http://www.strath.ac.uk/architecture
(3) ecilamar@unb.br, http://lecomp.fau.unb.br

II. Research Problem

These resources are being incorporated, though slowly, to architectural design and building construction in several countries [6]. These technologies are often referred to as CAD/CAM systems (Computer-aided Design / Computer-Aided Manufacturing) and also as CNC (Computer Numerical Control) processes [7].

The incorporation of digital fabrication to building design and construction processes is more slow and recent in Brazil than in other countries.

It has been popularly believed that such resources are not suitable to the reality of an emergent economy, such as the Brazilian one, because they would be more expensive than conventional processes. However, it has been already demonstrated by several authors that this does not result in a more costly architecture [5][8][9].

It has also been popularly believed that digital fabrication is yet not viable in Brazil due to the lack of CNC machinery in the construction industry. We will demonstrate in this paper that this is not true and that other reasons must be considered such as the AEC (Architecture, Engineering and Construction) education system. We will also show some initial results of teaching mass customization at early architectural education.

III. Hypothesis

We believe that the incipience of digital fabrication in the Brazilian AEC industry is not the result of a possible lack of technological resources in this country. It is important to stress that such resources are available and large in use by the Brazilian aircraft and car industries. It is even more important to stress that digital fabrication systems are already available, for example, in factories of steel structures and metal frames in this country [10].

We believe the present main reason for the incipient use of digital fabrication in building design and construction in Brazil is due to lack of information rather than to an alleged technological unavailability. This lack of information on its turn is due to our architectural education being still strongly based on the mass standardization paradigm. The weak connections between architectural schools and the

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_28, © Springer Science+Business Media B.V. 2010

construction industry in Brazil also contribute to the lack of information about fabrication processes available in this country.

Therefore, the objective of this paper is to demonstrate that the thesis of CNC technology unavailability as the main cause of it not being used in architecture is unsustainable. We will also show that some experiments in architectural education involving rapid prototyping and digital fabrication may provide a lead towards the solution of this problem.

IV. Research Method and Results

According to the 3rd Brazilian Inventory of Metal Forming and Cutting Machines [11] the total number of equipments in this area increased in 2.7% from 2006 to 2008. During the same period, the number of CNC machines increased in 44.4%. This suggests a technological potential in fast expansion which can be better explored.

A. CNC availability in Federal District construction industry

Most of the data in the aforementioned Inventory (96,5%) was gathered from the south and southeast regions of Brazil. In this paper we show the results of a survey with the construction industry in the specific region of the Brazilian Federal District. Companies of small, medium and large size and with different functions in the construction industry were surveyed.

Figure 1 shows the percentage of participation of different types of industries in the total of firms surveyed. Several companies were approached to establish a reasonably representative sample. Fourteen of those firms accepted to participate in the survey. The metal structure companies are the majority in the sample. This is due to the fact that this sector is the most dynamic of the construction industry in the process of incorporating CNC technologies. This choice was made under the assumption that this sector should influence others in the long term. As an example, the steel structure companies, through the manufacturing of metal molds, should influence concrete construction systems.

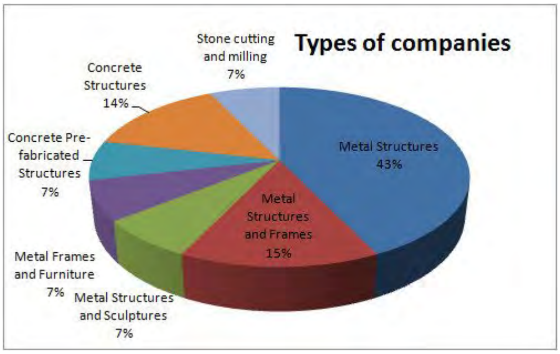


Figure 1. Types of companies surveyed.
(Source: authors of this paper)

Figure 2 below shows the percentages of companies that have only conventional technologies and those with CNC machines. The percentage of firms with CNC equipment is significant and supports our hypothesis. It is important to stress that 0.9 CNC machines per company in the surveyed geographic area is close to the national average detected by [11]. This result is particularly important considering that the Federal District in Brazil is an area predominantly administrative where only in the last 20 years significant industrial activities have emerged.

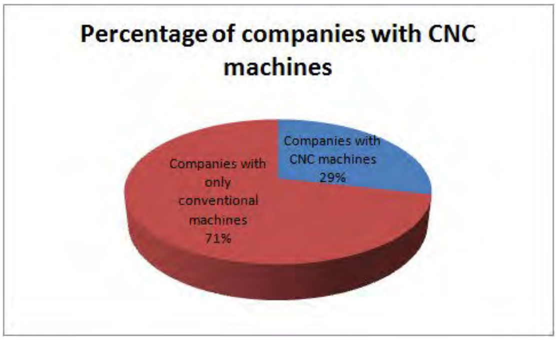


Figure 2. Percentage of companies with CNC technologies.
(Source: authors of this paper)

Figure 3 shows the percentage of companies that have CNC technologies taking into consideration only metal frames and structures.

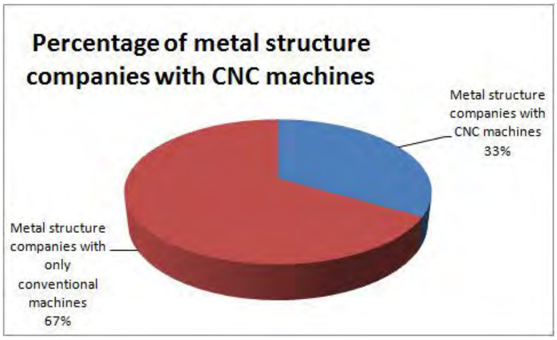


Figure 3. Companies of metal frames and structures with CNC technologies.
(Source: authors of this paper)

Figure 4 below shows the amount of CNC machines according the type of equipment. We use here a classification of four groups, i.e., 2D cutting, addition, subtraction and forming, in a similar way to that proposed by [12] and [13]. The 2D cutting CNC machines are the most common, but CNC forming (bending and stamping) is also significant.

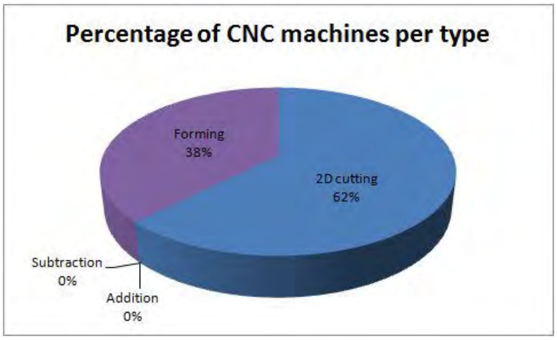


Figure 4. Percentage of CNC machines per type.
(Source: authors of this paper)

Figures 5 and 6 show oxyacetylene torches of 2D CNC machines cutting metal sheets at the *Mega Forte Metalurgica* and *Ferro Aço Badaruco* (http://www.badaruco.com.br), respectively, at the Federal District of Brazil.

Figure 5. 2D CNC cutting at Mega Forte, Brasília.
(Source: photo taken by authors of this paper)

Figure 6. 2D CNC cutting at Ferro e Aço Badaruco, Brasília.
(Source: photo taken by authors of this paper)

Figure 7 below shows a sculpture in steel sheets designed by the artist Darlan Rosa (http://darlanrosa.com/portu/jk.htm) and produced with CNC machines by the metallurgic metalúrgica *Ferro e Aço Badaruco* at the Federal District (http://www.badaruco.com.br). This example is important because it shows the same formal freedom that has already been explored in architectural design in other countries but remains unexplored in Brazil.

Figure 7. Sculpture of Darlan Rosa created using CNC technologies by Ferro e Aço Badaruco, Brasília
(Source: photo taken by authors of this paper)

B. CNC in Architectural Education in Federal District

It seems clear that the reason for the slow pace of CNC integration into the architectural design process in Brazil is not due to a lack of technological resources. The above mentioned survey has shown quite the contrary. The reason must reside elsewhere. We believe that the prevailing architectural education in our country must have an important role to play in this problem, to say the least.

Therefore, we have carried out several educational experiments at the Faculty of Architecture and Urban Design, University of Brasilia, Brazil, over the last two years. Those experiments were developed in the context of the second semester of the first year of the architectural design studio education. The focus of this semester is architectural language and expression with an emphasis on formal analysis and synthesis. The adopted design themes were an entrance porch and an exhibition pavilion.

One of our major objectives was to promote an understanding of the shifting paradigm in architectural design and construction at the beginning of this 21st century. We wanted the students to understand, as earlier as possible in their studies, the process of moving from mass standardization to mass customization in order to break free from the boundaries of orthogonal and serial architecture and to start exploring richer forms.

In order to achieve those goals we gradually introduced the paradigm of mass standardization (and its relation to the modern movement), the late 20th century paradigm of mass customization (and the benefits it may provide to the rising of new forms in architecture) and rapid prototyping.

The experiment was developed during four semesters and rapid prototyping was introduced at the end of the third one, on a limited range. Most students were requested to design using handmade scaled models. The design projects with more complex geometry were selected for rapid prototyping in a three-axe milling machine.

Figure 8 shows a rendered view of one of the students design project for the exhibition pavilion.

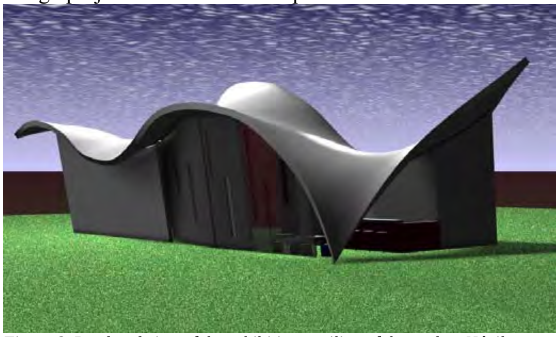


Figure 8. Rendered view of the exhibition pavilion of the student Núgila Ramos
(Rendered with formZ RenderZone Plus)

Figure 9 shows the prototype of roof of the design project shown above, just before being removed from the sustaining frame.

Figure 9. Rapid prototype of the exhibition pavilion of the student Núgila Ramos
(Milled in a Roland Modela MDX-40 three-axes milling machine)

The above example was one of design project specially selected for rapid prototyping.

We observed that the more theoretical background we gave to the students about digital fabrication and mass customization and the more information we provided about the existence of digital fabricators in our region, the more they felt confident to endeavor outside the boundaries of orthogonal and serial architecture.

The presence of CNC equipment in our school was important to convince the students of this technology's potential. However, one of the surprising results was that many students started designing taking into consideration the possibilities of the aforementioned paradigm shift even before we introduced rapid prototyping and they had started seeing its practical potential.

V. Conclusions

The results of our survey show that thesis of technological unavailability is not sustainable and that the presence of CNC technology in the Federal District of Brazil is already significant, although its distribution may be considered heterogeneous.

Our experiments with CNC at early architectural education show that the educational issue is a strong candidate for the problem of slow incorporation of digital fabrication and mass customization into architectural design practice in Brazil.

The implications are huge for the future of the Brazilian architecture considering the possibility of incorporating the mass customization paradigm and thus allowing for complex and non standardized components to be produced within reasonable economic constraints.

Acknowlements

We would like to thank the management of Gravia (http://www.gravia.com/), Mega Forte e Ferro e Aço Badaruco (http://www.badaruco.com.br) for kindly participating in survey described in this paper.

References

[1] Polette, Doug and Landers, Jack M. "Construction Systems", Goodheart-Willcox Company, Inc, Publishers, Illinois, 1995, pp. 45-49.

[2] Kieran, Stephen and Timberlake, James, "Refabricating Architecture", McGraw-Hill, New York, 2004, pp. 131-153.

[3] Kolarevic, Branko: Digital Production, in Branko Kolarevic, Architecture in the Digital Age – Design and Manufacturing, Taylor & Francis, New York, pp. 31-53, 2005.

[4] Schodek, D.; Bechthold, M.; Griggs, K.; Kao, K. M. & STeinberg, M. "Digital Design and Manufacturing – CAD/CAM Applications in Architecture and Design", John Wiley & Sons, New Jersey, 2005, pp. 339-344.

[5] Franken, Bernhard, "Real as Data", in in Branko Kolarevic, Architecture in the Digital Age – Design and Manufacturing, Taylor & Francis, New York, p. 138, 2005.

[6] Stacey, Michael; Beesley, Philip & Hui, Vincent, "Digital Fabricators", University of Waterloo School of Architecture Press, 2004, p. 6.

[7] Schodek, D.; Bechthold, M.; Griggs, K.; Kao, K. M. & STeinberg, M. "Digital Design and Manufacturing – CAD/CAM Applications in Architecture and Design", John Wiley & Sons, New Jersey, 2005, pp. 3-4, 237-296.

[8] Kolarevic, Branko: Digital Fabrication: Manufacturing Architecture in the Information Age, in ACADIA 2001, Buffalo, New York, 2001, p. 275.

[9] Schodek, D.; Bechthold, M.; Griggs, K.; Kao, K. M. & STeinberg, M. "Digital Design and Manufacturing – CAD/CAM Applications in Architecture and Design", John Wiley & Sons, New Jersey, 2005, p. 341.

[10] Cunha, Edgard L., Gonçalves, José R., Souza, José R. A., Silva, Sílvio P.(Editores): Corte e Conformação de Metais, Editora Aranda, São Paulo, 2008, número 44, pp. 70-91.

[11] Cunha, Edgard L., Gonçalves, José R., Souza, José R. A., Silva, Sílvio P.(Editores): Corte e Conformação de Metais, Editora Aranda, São Paulo, 2008, número 44, p. 71.

[12] Schodek, D.; Bechthold, M.; Griggs, K.; Kao, K. M. & STeinberg, M. "Digital Design and Manufacturing – CAD/CAM Applications in Architecture and Design", John Wiley & Sons, New Jersey, 2005, pp. 3-4, pp. 239-244.

[13] Kolarevic, Branko: Digital Fabrication: Manufacturing Architecture in the Information Age, in ACADIA 2001, Buffalo, New York, 2001, pp. 268-277.

Application of Project Based Study in the Learning Process: A Case Study

Ashfaque A Chowdhury, Mohammad G Rasul and M. Masud K. Khan

Abstract— **The study is a step toward studying the advanced technology specially to learn the control system of CNC lathe machine. This paper describes the development of a closed loop control system with encoder mounting as real-time control device for a conventional lathe. A personal computer is used for calculation as well as for controlling the motion of the drives. The servomotors are used for controlling the movement of carriage. A program is developed using Turbo C for evaluating the overall system. The developed control system is applied to evaluate work piece dimension in real time with the integration of production process. Encoders are used to collect real-time machining data. It is found that the errors are less than 0.005 mm of turning example part on X and Z axes. The average error is approximately 0.002 mm.**

Index Terms— **CNC, Lathe, real-time control system.**

I. Introduction

LEARNING document methods and supporting informational materials in the higher education can be improved for varying disciplines by engaging students to unravel real life problems through project based investigation. The aim of engineering education is to develop skilled manpower to solve real life engineering problems and to meet society's needs with advances in knowledge. Gray [1] investigated the integration of case study based investigation in engineering courses to establish a practice where undergraduate engineering students exercise their growing technical knowledge and analytical capability in the context of real engineering problem. The applied principals of technical investigations were generic in nature. Harris [2] demonstrated the implementation of the project based design methodologies in an undergraduate course by integrating the relevant theories with hands on design projects. It not only benefits the students but also changes the instructors to integrate theory into practice. Holt [3] reported that students often receive little practice in closing the gap between engineering necessity and technical knowledge. This lacking may lead to mistakes by the students in treating the engineering problems in practice. Jennings [4] reported on a case where students were asked to expand on the professional dimensions of the implications of the problems they solved as a part of class room teaching and the outcomes of the solutions to engineering problems.

Navaz et al. [5] introduced a new approach in teaching Thermal/Fluid Sciences to undergraduate students through inclusion of new and advanced technologies into the curriculum using latest computational and experimental methods. Musto [6] described the outcomes of the introduction of a project based approach in a numerical methods course where the subject concepts were related to a term-long design project. Student teams were required to report on the design methodology and the results through formal design reports and a presentation. The integrated project based approach assisted the students to enhance their enthusiasm and understanding of the course materials [6].

Mustoe and Croft [7] reported the increased emphasis on open ended design exercises and case studies in engineering curricula. Gomes [8] adopted case studies in modern engineering teachings to improve integration of cross disciplinary education and to develop a professional outlook in the young engineers. Feisel and Rosa [9] identified the lack of coherent learning objectives in engineering laboratory exercises and demonstrated the impact of this shortcoming in limiting the effectiveness of these exercises. Feisel and Rosa [9] encouraged engineering students to seek knowledge beyond the classroom theories. DeBartolo and Robinson [10] demonstrated the successful implementation of design and experimentation in engineering curriculum to develop the students' concepts on design, build and prove the functionality of the designed project.

The integration of project based investigation into the fluid mechanics course described in this paper has not only provided the students an insight into the fundamentals of fluid mechanics, but also developed students' abilities to carry out professional tasks in real life. The study has also reported that the suitability of project based study along with conventional lecture classes in fluid mechanics course.

II. Summary of the Course

In the Department of Sustainability and Infrastructure at CQUniversity, teaching of fluid mechanics course is offered through dedicated course work, laboratory experiments and group design experiments. The core components of the course consists of a collection of weekly modules and contains the general characteristics of fluids, manometry, fluid statics,

Manuscript received November 2, 2009. This work was supported by the Department of Mechanical and Chemical Engineering, Islamic University of Technology.

A. A. Chowdhury is with the CQ University, Rockhampton, Qld 4701, Australia (corresponding author phone: +61 4 1176 7886; fax: +61 7 4930 9382; e-mail: a.chowdhury@cqu.edu.au).

M. G. Rasul is with CQ University, Rockhampton, Qld 4701, Australia (e-mail: m.rasul@cqu.edu.au).

M M K Khan is with CQ University, Rockhampton, Qld 4701, Australia (e-mail: m.khan@cqu.edu.au)

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_29, © Springer Science+Business Media B.V. 2010

analysis of incompressible flow in pipes, analysis of buoyancy and stability of floating bodies and measurement of fluid flow. It introduced the methods of analysing fluid systems using the conservation of mass and momentum equations, Navier Stoke's equation combined with the concept of a control volume. To solve the problems in fluid mechanics course, students familiarised with the theories of incompressible flows in pipe systems, Euler's equation, Bernoulli's equation and the use of similitude and modelling principles and techniques.

Students were required to act professionally in presenting information, communicating, working and learning individually and in peer learning teams to achieve the learning outcome of the fluid mechanics course. Initially the fluid mechanics course was offered to on-campus students. Later, the existing course was adapted for Interactive System-wide Learning (ISL) delivery on other regional campuses. An online lecture delivery was developed for flexible course offerings with a one week on-campus mandatory residential school requirement. The restructured course materials can be used both as supporting teaching materials and as reference materials in project work for research and self study.

III. Students learning activities

At the beginning of the course, all students were given an overview of the course to make clear the expectations about how the course would operate. Flex students were introduced to online learning practice and were assessed on online learning skills required for this program. Students were encouraged to study and respond to the review questions and solve exercise problems. They were also familiarised with the real fluid systems and flow phenomena through the laboratory exercises. Students were required to study the text and other course materials as a guide to solve the workbook problems, virtual laboratory activities and other assessment items.

Students were able to discuss their difficulties with problems in the discussion list or email the course lecturer. Students were expected to follow a self-directed study schedule which met the required deadline. The weekly schedule of activities for on campus students were Lectures (two hours per week), Tutorials (two hours per week), Laboratory (two hours per week). A residential school was conducted for flexible and other regional campus students after the midterm vacation. Tele-tutorials over Skype and/or local tutorials were arranged for off-campus students.

IV. Addressing of graduate attributes

In the design and development process of a technical course for the engineering students, there are several objectives need to consider such as the goals, resources, course learning outcome, and last but not the least the graduate attributes. The targeted graduate attributes were Science and Engineering knowledge, Effective communications, Technical competency, Problem solving and systems approach, Functionality in a team, Social, cultural, global and environmental awareness, knowledge on sustainable design and development, professionalism and ethics and lifelong learning.

The graduate attributes mapping was prepared to obtain an indication of how the fluid mechanics course contributed to achieve course learning outcomes and graduate attributes and to understand better how attributes are developed as students' progress from one year to next year. This is done by indicating the graduate attributes targeted by each course learning outcome. The indicative marks are placed under the headings of Teaching, Practice and Assessment to illustrate the course emphasis (low/medium/high).

V. Assessments

To pass the course students had to attempt all the assignments, laboratory reports and complete all examinations and obtain a grade of at least 50% and a passing grade in the formal exam. Assessments for the course were on the basis of laboratory reports (20% weighting), a fluid mechanics project plan (20% weighting) and investigation (20% weighting), workbook submission (pass/fail in competency) and formal examination (40% weighting).

A. Laboratory Exercises

There were three components in laboratory exercises – a formal lab report, video lab questions and virtual lab design and experimentation. The first two components were group assessment items and each group were required to submit a report on the assessments. Every group member contributed to the conduct, preparation and write up of the laboratory report. The objective of involving students in video lab questions and virtual lab activities is to develop higher order thinking from the problem statements and thus a deep understanding of the knowledge on the subject matters. Practical videos in most chapters of weekly schedules were uploaded on the course delivery site.

The students were asked to solve the related practical questions demonstrating the steps with related background knowledge for a virtual lab problem. In the design and experimentation component, the students were encouraged to design an experiment in fluid mechanics from any of the topics covered in the weekly modules that the students would like to improve further in future. Students were required to submit a report on the selected experimentation covering the necessity for the investigation, a description of the work, the parameters requiring observation and the safety procedures. Students were also required to complete the entire laboratory exercise including the drawing of graphs and final answers on issues such as calibration of obstruction meter, flow in pipes, stability of a floating body and velocity profile and pressure losses in a smooth pipe.

B. Engineering Fluids Project

The aim of an engineering fluids project is to ensure that the future engineers will comply with the principles of sustainability though improving their research ability and knowledge of fluid mechanics. The projects related to basic principles of fluid mechanics assist the students to enhance their learning experiences and to achieve the graduate attributes through a team based design approach utilising the concept of sustainability to meet the needs of present and future generations. The students experienced a productive pedagogy in fluid mechanics through improved knowledge & understanding.

Several project options were available for the students. In a project proposal, on part of land assumed, the right side was assumed to construct an engineering complex. Based on the knowledge in the course, a group of students were responsible to design the supply of drinking water and the management of waste water facilities for the engineering workshop. Another two groups of students were assigned to design the same for a shopping complex and a cinema theatre. There were several steps suggested in the course profile for the students to follow and report to the campus tutors on the progress of the project.

In the first two to four weeks, students were required to define problems and project scope through the fundamentals of fluid mechanics learnt, and then proposes a sustainable concept design complying with course learning outcomes. In week five, students were required to present the progress in groups identifying the contribution of each member. In the next three to four weeks, students prepared the preliminary sustainable design with specifications. Often students revisited the preliminary plan and refined the concept design with the feedback from course tutors and lecturer. Later in the term, students finalised and documented the sustainable design with a strategy to implement in future. It was required to build a physical prototype design to have a nice memory of the young engineers' efforts in future.

The amount of traditional homework problems assigned was reduced approximately by half. The design project was assigned to groups of four or five students at the start of the course. Handing out the project immediately at the outset of the course, where students are largely unfamiliar with the material required for the completion of the project, renders the learning process goal driven. This approach is in support of the life-long learning scenario for which students ought to be prepared and where the learning typically occurs on a needs basis in an interactive and often collaborative learning mode.

The submissions of two written progress reports were required in order to guide students through the design tasks and enforce progress throughout the entire semester. In addition, a total of two class periods throughout the semester were allotted for a progress presentation and a final presentation by each student team. It was found that the project based learning concepts enhanced students' learning experiences and assisted in the achievement of key attributes. The students experienced a dynamic pedagogy in fluid mechanics through improved content quality, connectedness to real world problems, a supportive team learning environment and recognition of differences between the team members.

C. Other Assessment Items

To improve the research capability and gain knowledge and improve understanding in fluid mechanics, students were asked to prepare a teaching and learning journal in fluid mechanics through the lectures, tutorials, experiments and project work. Students were expected to solve higher order problems and open answers which require researching on the weekly topics. The workbook contributed to the overall grade of the course as it supplemented the assignment work and confirmed students overall competency. They were required to show the application and understanding of the subject matter and to use the correct procedures and calculations in their workbook answers. Attainments of competency attracted extra marks to move to a higher grade when short by a few marks. Students were also required to sit for closed book final exam.

VI. Assessments Criteria

Engineering fluid projects and workbooks were evaluated based on the evidence of the application and integration of knowledge and understanding of the material studied and communication and justification of the subject matters. It was kept in consideration that students demonstrated accurate use of engineering terms, collection and organisation of information in the various forms of presentation. It includes well structured solutions, use of engineering mathematical reasoning and proof to develop logical arguments, recognition of the effects of assumptions used evaluation of the validity of arguments and justification of procedures.

The laboratory components of the assessment items were assessed based on reporting of key elements required to undertake the laboratory sessions, clarity of expressions, appropriate referencing of sources, accurate use of fluid mechanics knowledge, presentation of mathematical formulation, clarity and logical presentation of ideas and arguments by means of data analysis and synthesis. To evaluate the ability of the students on the application and integration of knowledge and understanding - evidence of interpretation, analysis, identification of assumptions, selection and use of symbols of strategies and procedures while showing initiative to solve the problem, were considered. In some cases, it was found that the problem has not been solved, however substantial progress towards a rational solution and shows evidence of appropriate interpretation, analysis, identification of assumptions and selection of strategies and procedures. Additionally, the problem has been solved however, there is a lack of appropriate evidence was also reported.

VII. Students performance

Course history showed that, after introduction of the integration of project based investigation in fluid mechanics course, on average (based on three years data), about 4.6% of the total enrolled students withdrew from the course after enrolment and the failing rate dropped to 6.5%. Most of the students performed very well and their overall performance was quite impressive. On an average, 28.7% students received high distinction and 30.6% students received distinction in the course. As shown in Fig 1, the comparisons of student performances over 2007 and 2008 demonstrate that the failing rate of the students in the course was reduced from 14.3 percent to 5.4 percent. The blended assessment mechanism incorporating lab components, project works, workbook and final exam assisted students to understand the subject matter and to develop a higher level of skills in an advanced engineering topic like fluid mechanics.

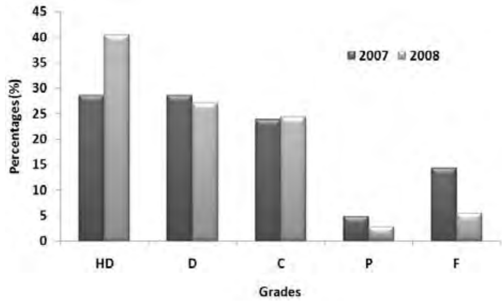


Fig. 1 Indication of students' performance in the course

VIII. Further developments

Integration of project based study based on practical engineering problems in mechanical engineering courses is required to ensure the professional and engineering context of applied science and to develop subsequent learning methods. The program review committee meeting at CQUniversity suggested the Program Director the course outcome, students' feedback, and peer assessments (tutors and course coordinator) for potential improvement of the next offering of the course. In 2009, based on student comments and peer assessment, the course was progressively restructured to maintain the relevance of the content and satisfy the demand and need of the industry and community. Dr Philip Morgan's (the University of Newcastle) flowchart (Fig 2) was followed to incorporate the feedback from the past students, tutors, lecturer and course coordinator. To collect the feedback received from the student, course evaluation and focus group interviews were conducted. In course evaluation the students where asked about the active learning (actively engaged and participated in collaborative learning), assessment (satisfactorily demonstration of the acquisition of the knowledge, skills and attitudes needed), expectation (to learn to satisfactorily complete this course materials), relevance of the course materials (satisfactory development the appropriate knowledge, skills and attitudes to future professional career) and overall satisfaction (positive learning experience). Some feedbacks were received from open ended comments by the students in the emails. The summary of the results of the feedback and recommendations have been articulated in the course improvement flowchart. In has been reported that the engineering projects relevant to the theories not only solve professional cases but also encourage the students to involve in an encouraging manner.

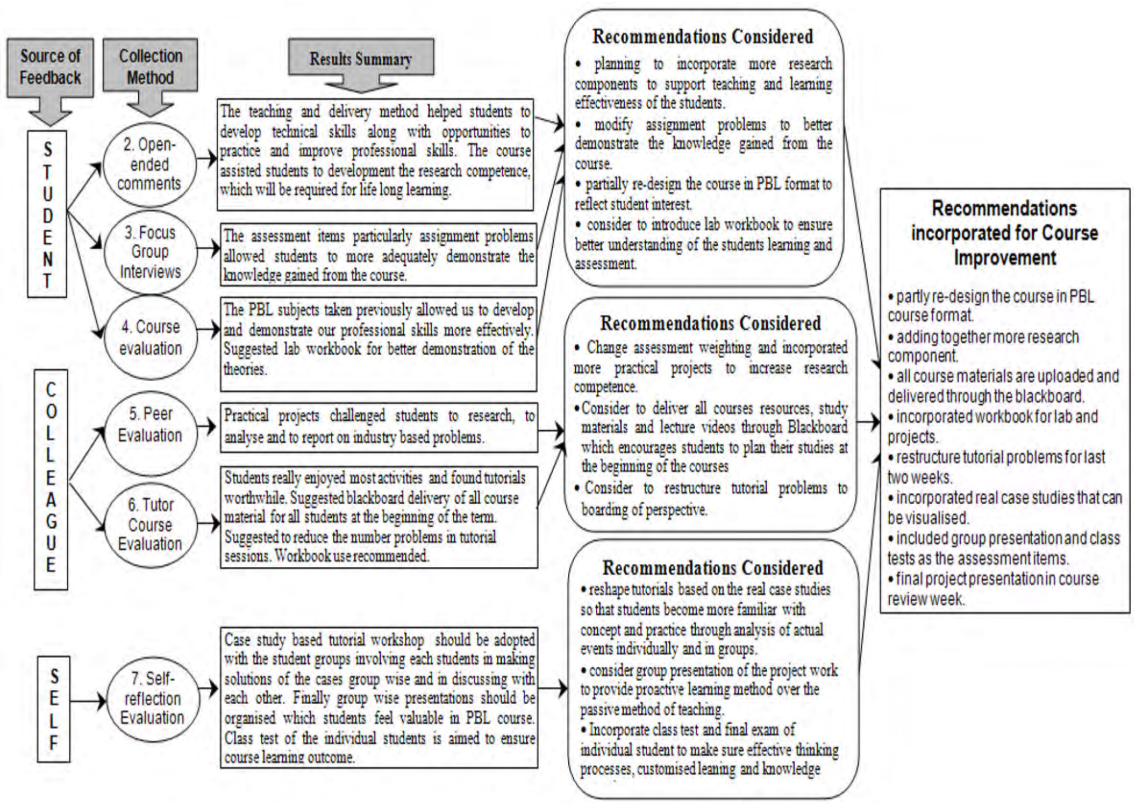


Fig. 2 Flowchart of course improvement process

IX. Concluding Remarks

It is intended to teach the fluid mechanics course by creating a learning centred environment which will encourage the students for creative and critical thinking and lifelong learning. It is necessary to prepare undergraduate engineering students to design, and analyse real life problems. The paper describes an integrated assessment methodology which was implemented in a fluid mechanics course and demonstrates its effectiveness in facilitating engineering undergraduate students to become efficient problem solvers of real life problems. The success of the assessment procedures were evaluated based on the successful completion of the course, students' grades, students' withdrawal rate from the course and the students' overall perception of benefits of the course studied. Integrated assessment methods assisted with strengthening previous and prerequisite learning requirements. Therefore, benefits the subsequent learning process is achieve through excellence in research and teaching. The assessment items offered students experience of real life problems in a professional manner. Further improvement of the course is in progress considering the feedback from all the stakeholders of the course.

Acknowledgment

The authors gratefully acknowledge the contributions of Dr Ahmed Elgezawy and Mr Fons Nouwens who generously assisted in the development and implementation of the course in the faculty. Special appreciation goes to Mr Tim Mcsweeney for his remark to improve the manuscript. The authors also acknowledge the support of the Learning and Teaching Education Research Centre (Special Interest Group – Future Engineering Education Directions), which is established and supported under CQUniversity's Research Centres Program.

References

[1] T. Gray, "Integration of case study technical investigations in honours/masters engineering courses," *International Journal of Mechanical Engineering Education*, vol. 34, pp. 121-130, 2006.

[2] D. Harris, "A Case for Project Based Design Education," *International Journal of Engineering Education*, vol. 17, pp. 367-369, 2001.

[3] J. E. Holt, "On the nature of mechanical engineering - an engineering ethos," *International Journal of Engineering Education*, vol. 12, pp. 332-338, 1996.

[4] A. Jennings, "Viewpoint: the need to upgrade experimental learning," *International Journal of Engineering Education*, vol. 14, pp. 83-88, 1998.

[5] H. K. Navaz, B. S. Henderson, R. M. Berg, and S. M. Nekcoei, "A new approach to teaching undergradaute thermal/fluid sciences - courses in applied computational fluid dynamics and compressible flow," *International Journal of Mechanical Engineering Education*, vol. 30, pp. 35-49, 2002.

[6] J. C. Musto, "A project based approach to teaching numerical methods," *International Journal of Mechanical Engineering Education*, vol. 30, pp. 233-247, 2002.

[7] L. R., Mustoe, and A. C. Croft, "Motivating engineering students by using modern case studies," *International Journal of Engineering Education*, vol. 15, pp. 469-476, 1999.

[8] V. Gomes, "Consolidation of engineering education through industrial case studies," *International Journal of Engineering Education*, vol. 18, pp. 479-484, 2002.

[9] L. D. Feisel, and A. J. Rosa, "The role of the laboratory in undergradaute engineering education," *Journal of Engineering Education*, vol. , pp. 121 -130, 2005.

[10] E. DeBartolo, and R. Robinson, "A freshman engineering curriculum integrating design and experimentation," *International Journal of Mechanical Engineering Education*, vol. 35, pp. 91-107, 2007.

Ashfaque Ahmed Chowdhury obtained his B.Sc. in Mechanical Engineering degree from Islamic University of Technology, Bangladesh in 2003. In 2008, he completd Master of Engineering degree at Central Queensland University, Australia. Since 2004, he worked as a lecturer in the Department of Mechanical and Chemical Engineering at Islamic University of Technology and in the College of Engineering and Built Environment, Central Queensland University. His current research interest include intelligent control system, manufacturing technologies, computer numeric control manufacturing, computer-aided design and manufacturing, building energy simulation, energy conservation and management, low energy cooling technologies.

M G Rasul is a Senior Lecturer of Mechanical Engineering at the College of Engineering and Built Environment, Central Queensland University, Australia. He received his PhD and MEng degree from The University of Queensland, Australia and Asian Institute of Technology, Thailand respectively. He is specialized and an expert in energy technologies, process industries, thermo-fluid engineering including combustion and process engineering, and light metals. He has published over 100 research articles in reputed journals and conferences on a wide range of research interests.

M. Masud K. Khan is an Associate Professor and Head of Department of Infrastructures in the Faculty of Sciences, Engineering and Health, Central Queensland University, Rockhampton, Queensland 4702, Australia. He received his MS (Mech) with 1st class Honours from Moscow Institute of Petrochemical and Gas Industry in 1982. Subsequently he worked with the oil industry for 2 years. He obtained his PhD in engineering from the University of Sydney in 1990. His teaching, research and consulting activities are in the area of non-Newtonian fluid mechanics, thermofluid engineering and rheology of industrial fluids and polymeric materials. He has authored more than 95 refereed journal, conference and technical papers. He is a member of American Society of Rheology, Australian Society of Rheology, Engineers Australia and the Australasian Association for Engineering Education (AAEE).

Interdisciplinary Automation and Control in a Programmable Logic Controller (PLC) Laboratory

Jack Toporovsky, Christine Hempowicz, and Tarek M. Sobh
School of Engineering, University of Bridgeport
ytoporov@bridgeport.edu, chemp@bridgeport.edu, and sobh@bridgeport.edu

***Abstract* – The University of Bridgeport's School of Engineering created a new PLC Industrial Control Lab to accommodate lab-based, hands-on training using different types of Programmable Logic Controllers (PLCs). The Lab and associated courses were designed to meet ABET's EC2000 standards and prepare students for the workforce.**

I. Introduction/Background

At the beginning of the new millennium, engineering education programs were challenged to meet the latest set of standards established by ABET, known as *Engineering Criteria 2000* (EC2000). EC2000 emphasized outputs, as expressed in 11 learning outcomes [1]. In particular, upon program completion, graduates should have acquired the knowledge and skills to solve problems, using techniques and tools of engineering practice in a collaborative setting.

This approach to engineering education is not entirely new. Early engineering education focused on skill acquisition through apprenticeships, practical training, and the like, until the end of World War I [2]. The emphasis on engineering science and mathematics in engineering curricula was not introduced until the mid-1950s. This had the effect of producing engineers who could analyze a problem but were less prepared to design and apply solutions [2]. EC2000 was, in part, a response to the need for engineering graduates to be able to seamlessly transfer classroom and laboratory knowledge and practice to the workplace. Engineering graduates must be "job ready" and practiced in working as part of an interdisciplinary team [3] and [4].

II. Problem Statement

The University of Bridgeport's School of Engineering (UBSOE) identified the need for a Programmable Logic Controller (PLC) Industrial Control Lab to support laboratory-based courses in which students could be educated in technological processes that are part of the field of automation control. Course curricula should support student outcomes outlined in EC2000's Criterion 3. Program Outcomes and Assessment, specifically 3a-e, j, and k. Further, course curricula should be interdisciplinary such that electrical engineering and mechanical engineering students could enroll in the same course. Finally, students should have the opportunity to learn on leading PLC models in order to be best prepared to work in the private sector. The lab should be equipped, then, with Allen-Bradley and Mitsubishi PLCs since the former is widely used in the U.S. and the latter is widely used throughout Asia.

III. Methodology

UBSOE's existing Interdisciplinary Robotics, Intelligent Sensing, and Control (RISC) Laboratory was expanded to incorporate the new PLC Industrial Control Lab. The new PLC lab was equipped with electrical cabinets and control panels outfitted with five Allen-Bradley and four Mitsubishi PLCs [5]-[12].

Two courses were developed: EE 463: Industrial Controls and EE 464: Programmable Logical Controls.

In EE 463, students learn basic concepts in pneumatic controls and different types of sensors

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_30,

used widely in industry, robotics, and vision concepts.

In EE 464, students learn basic and advanced concepts and applications for PLC programming. In addition, students learn to integrate various components with PLCs, such as sensors, switches, vision cameras, variable speed drives, etc.

Instructional design for each course was more focused on inquiry and problem-based learning within the context of classroom community, and less focused on lecture-style instruction. All are supported in the literature [13] and [14].

A. Equipment Design and Integration

In the PLC Industrial Control Lab, all control cabinets share the same local network with different types of PLCs and Human Machine Interfaces (HMIs), so students are able to communicate to any of them from the class computer stations.

A team of PLC students designed an integration of different brands of control components into one application. The components are: SLC5/03 Allen-Bradley PLC, Network Card 1761-NET-ENI, Mitsubishi Variable Frequency Drive, and Cutler Hammer PM1700 Touch Screen. For this project, students were involved in wiring, communications cable design, equipment programming, and troubleshooting.

Students produced the layout design and wiring diagrams for the project to integrate the SLC5/03 PLC with the Mitsubishi Variable Frequency Drive (VFD) and Cutler Hammer Touch Screen, and to establish a network to the PLC. The design layouts and wiring diagrams were created using AutoCAD. Students were able to build a control cabinet with output to an AC motor, programming of the PLC, VFD, touch screen, etc. (See Figure 1.)

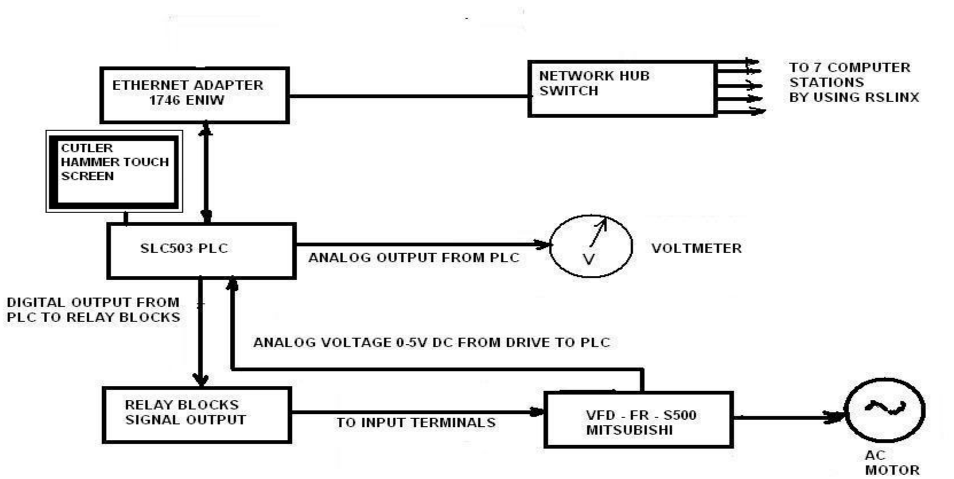


Fig. 1. Any computer station can establish communication to the SLC5/03 PLC through the network hub switch and through the internet adapter (# 1746ENIW). At the same time, an AC motor (3-phase AC) runs commands from the PLC, transferring from the relay module to the Mitsubishi Variable Frequency Drive (VFD/FR/S500). The value of the frequency is converted to an analog voltage (0 to 5 volts) from the VFD; it is then read by an analog card. In this way, a closed-loop communication link is established between the PLC and the VFD.

During the project, students learned how to use the software FR Configurator to run the motor with different parameter sets and observe the effects. After the students decided which parameters were needed for the operation, they downloaded them into the VFD.

At the end of the project, different touch screen recipes for various applications had been created.

RECEIPE 1

Selector switch # 1 should be at the ON position. Press the start push button on the touch screen and select a value for the motor's speed. (The motor will run at one of three predetermined speeds.) The motor speed control parameter (expressed as voltage) is shown on the touch screen. Also shown is the motor's speed in RPM's.

RECEIPE 2

Selector switch # 1 should in the OFF position. In this mode, the motor will run continuously, first at a low speed (typically 50 to 100 RPM's), then at a medium speed (approximately 100 to 300 RPM's), then at a high speed (300 to 500 RPM's), then back to low speed, medium speed, etc. The analog voltage drive speed and RPM are displayed on the touch screen.

RECIPE 3

This mode is designed to control the jogging operation by pressing the jog-forward or jog-reverse pushbutton or touching the jog-forward/reverse icon on the touch screen. The motor will rotate forward or reverse, depending on which jog button is pressed. The analog voltage drive speed and RPM are displayed on the touch screen.

RECEIPE 4

Establish motor control by reading the value potentiometer on the front panel. The speed can be increased from 0 RPM to a maximum of 500 RPM (as controlled by the PLC) by turning the potentiometer knob clockwise. To perform this test, the VFD parameter # 79 should have a value of 4.

B. PLC Programming and Robot Integration

The first author worked with a team of PLC students to design an automation conveyor line to implement closed-loop communications between the Mitsubishi robotic arm Roboware™ software and FX series Mitsubishi PLC. See Figure 2.

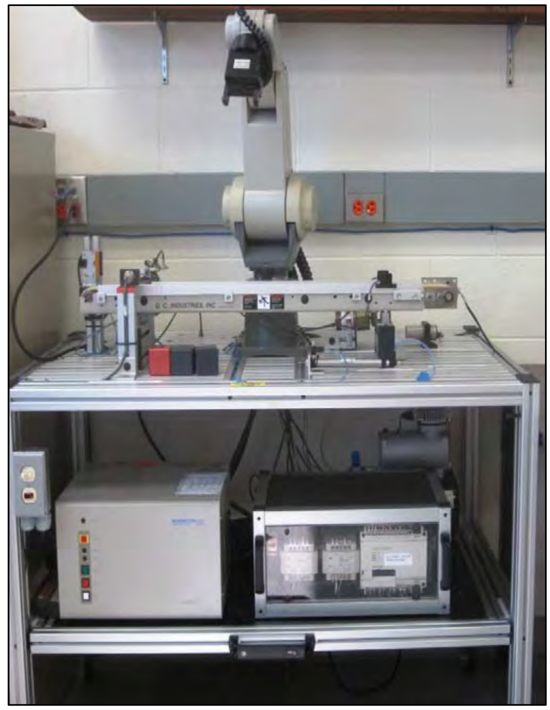


Fig. 2. Robotic arm is shown, controlling a conveyor line, with optical, proximity, and capacitive sensors. A Mitsubishi PLC is located on a shelf below the arm and to the right, and a Mitsubishi RV-M1 robotic unit is on the same shelf, to the left.

The major task of the project was to establish the communication and movements of the Robotics Arm using the Mitsubishi Movemaster RV-M1 series robotics kit. When the photo eye sensor detects the metal or plastic object on the loading area, the pneumatic cylinder then pushes the object to the part-detecting sensor (proximity sensor) (See Figure 3). If the object is plastic, the conveyor starts and stops at "position one," which is closer to the sensor. The robotic arm will move to pick up the plastic object and place it into "box one". If the object is metal, the conveyor stops when it is detected by the track sensor (see-through eye sensor) on the conveyor belt. The robotic arm moves to that position to pick up the object and place into "box two".

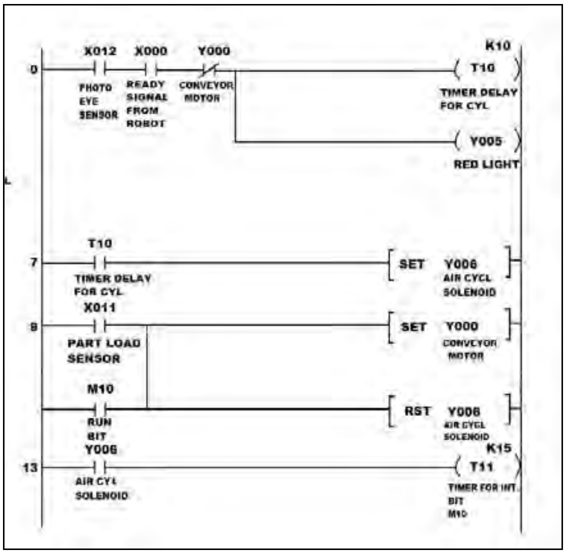


Fig. 3. Controlling the belt movements (Y6), part present (X12), red light indicator (Y5) and ready signal from robot unit (X0), etc, presented in the ladder diagram. Constant communication between the robot controller and Mitsubishi PLC is provided through executing this program and running Roboware™ software.

The inputs and outputs are on the conveyor line, operated using the GX-Developer (Mitsubishi software). See Table 1.

TABLE 1

PROGRAMMING OF PLC AND ROBOT

Input	Address	Output	Address
Capacitive Proximity Sensor	X10	Conveyor Motor	Y00
Part Load Sensor (prox. sensor)	X11	Gear Motor	Y01
Photo Eye Sensor (loading area)	X12	Red Light	Y05
Track Sensor (end of the conveyor)	X13	Air Cylinder	Y06

The command line to enter some commands such as MJ (Move Joint) is as follows:

Function: Turns each joint the specified angle from the current position. (Articulated interpolation)

(1) The least input increment of the turning angle is 0.1, e.g. specify 15.2 for 15.2.

(2) The open/close state of the hand does not change before and after the movement. Error mode II occurs before the axis motion if any turning angle entry exceeds the robot operational space.

(3) The default turning angle is 0.

The position data table for pick and place scenario loaded into robotic amplifier can be found in Table 2.

TABLE 2

Robotic Arm Position Data (PD)					
1	+254.6	+427.2	+174.6	-46.9	+32.2
4	-259.0	+185.1	+212.0	-87.0	+32.2
5	-256.6	+187.2	+236.0	-87.0	+32.2
6	-259.0	+185.1	+199.0	-87.0	+32.2
7	-259.0	+185.1	+235.0	-87.0	+32.2
8	-259.0	+188.1	+192.0	-87.0	+32.2
9	-259.0	+185.1	+263.0	-87.0	+32.2
10	-317.4	-24.5	+263.0	-86.9	+32.2
11	-324.5	-25.1	+43.1	-89.0	+32.2
12	-324.5	-25.1	+36.1	-89.0	+32.2
13	-324.5	-25.1	+44.1	-89.0	+32.2
14	+299.1	+182.9	+233.8	-81.0	+145.1
15	+297.9	+179.1	+196.4	-86.0	+145.1
16	+302.5	+184.8	+258.3	-76.9	+145.1
17	+22.2	+353.8	+258.3	-76.9	+174.7
18	+23.3	+371.4	+58.8	-92.0	+174.7
19	+28.8	+397.2	+44.9	-87.0	+174.7
20	-14.4	+281.1	+515.6	-49.0	+179.9

C. *Assessment*

Assessment of student understanding and ability to implement a PLC's interaction with sensors, motors, and other objects were performed for both short-term and long-term assessment. (1) Short-term assessment took place within the timeframe of an individual course. Each student was assessed by performance on eight homework assignments (20 percent of course grade), an in-class mid-term exam (30 percent of course grade) and an individual, open-ended project, in which each student was expected to incorporate codes, commands, and design functions (50 percent of course grade). (2) Long-term assessment is conducted after course completion and post-graduation. Each student is surveyed periodically to determine how each component of the lab course has helped the graduate in his or her professional employment and career.

IV. RESULTS

Students who successfully complete a laboratory course are able to: understand and implement different types of software packages for the purpose of robot control (such as GX-Developer or Roboware™ for Mitsubishi products or RS-500 and RS-5000 and Factory-Talk for Allen-Bradley products); and program Human Machine Interface (HMI) screens to integrate them into PLC control automation lines. After graduation, they can perform the tasks listed above and communicate intelligently with members of a team of designers working on automating any industrial process.

V. CONCLUSIONS AND FUTURE WORK

UBSOE's PLC Industrial Control Lab will continue to provide critical laboratory education for graduate engineering students. Further, the authors envision that the Lab will assist local companies in hardware design and manufacturing and provide the necessary software programming for different applications in process automation.

ACKNOWLEDGEMENT

The authors acknowledge the equipment contributions by Gibson Engineering Company, Incorporated of Norwood, MA and HESCO of Rocky Hill, CT.

REFERENCES

[1] L.R. Lattuca, P.T. Terenzini, and J.F. Volkwein. (March, 2006). *Engineering Change: A Study of the Impact of EC2000*. ABET, Inc. Baltimore, MD. [Online]. Available: http://www.abet.org/Linked%20Documents-UPDATE/White%20Papers/Engineering%20Change.pdf

[2] P.C. Wankat, R.M. Felder, K.A. Smith, and F.S. Oreovicz, "The scholarship of teaching and learning in engineering," in *Disciplinary Styles in the scholarship of teaching and learning: Exploring common ground,* M.T. Huber and S. Morreale, Eds., Washington, DC: AAHE/ Carnegie Foundation for the Advancement of Teaching, 2002, pp. Available: http://www4.ncsu.edu/unity/lockers/users/f/felder/public/Papers/Scholarship_chapter.pdf

[3] D. Richter, "Infusing an interdisciplinary automation experience in engineering technology education," in *Proceedings of the 2007 American Society for Engineering Education Annual Conference & Exposition*, Honolulu, HI.

[4] T.J. Stiller and G.R. Johnson, "Constituent influences on engineering curricula," in *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*, Salt Lake City, UT.

[5] C. Mirman, "Educating future engineers requires teamwork," *Automation Notebook*, Fall Issue 12, 2008, [Online]. Available: http://automationnotebook.com/2008_issue_12/featurestory_issue12_2008.html

[6] *E-Designer for the E1000-series Reference Manual,* Mitsubishi Electric, MA 00759A, 2005-03.

[7] *MicroLogix 1100 Programmable Controllers, Bulletin 1763, Instruction Set Reference Manual,* Rockwell Automation, Inc., Publication 1763-RM001B-EN-P, April 2007.

[8] *FX Series Programmable Controllers Programming Manual,* Mitsubishi Electric, Manual number JY992D88101, Manual revision D, April 2003.

[9] *Roboware ™ Programming/Control Software for Mitsubishi Robots Ver 2.xx User Manual*, KakeWare©, RV-M1 & RV-M2 Models.

[10] *Melsoft FR Configurator Instruction Manual VFD Software*, Mitsubishi Electric, FR-SW2-SETUP-WE.

[11] *Logix 5000 Controllers Common Procedures Programming Manual*, Rockwell Automation, Inc., Publication 1756-PM001I-EN-P, January 2007.

[12] *SLC 500 Instruction Set Reference Manual*, Rockwell Automation, Inc., Publication 1747-RM011E-EN-P, January 2006.

[13] K.C. Bower, T.W. Mays, and C.M. Miller, "Small group, self-directed problem based learning development in a traditional engineering program," in *34th ASEE/IEEE*

Frontiers in Education Conference, Savannah, GA, 2004.

[14] S.M. Lord and M.M. Camacho, “Effective teaching practices: Preliminary analysis of engineering educators, in *37th ASEE/IEEE Frontiers in Education Conference*, Milwaukee, WI, 2007.

Stereo Spectral Imaging System for Plant Health Characterization

S. C. Yoon[1] and C. N. Thai
University of Georgia
Biological and Agricultural Engineering Department
Athens GA 30602-4435
E-mail: thai@engr.uga.edu
Web site: http://www.engr.uga.edu/~mvteachr
[1]S. C. Yoon is now at USDA, ARS, Richard B. Russell Research Center, Athens, GA 30604-5677, seungchul.yoon@ars.usda.gov

***Abstract* - Three-dimensional (3D) measurements of whole plants may provide detailed structure information about plant growth patterns and also complement existing X-ray systems for the below-ground part. In addition to this structural characterization of plants, spectral information may also biochemically characterize plants' health. A stereo vision technique is a cost-effective and rapid imaging technique for measuring and reconstructing 3D structures. The Normalized Difference Vegetation Index (NDVI) requiring measurements of two spectral wavelengths in the NIR and red spectral regions has been widely used in remote sensing as an index to estimate various vegetation properties including chlorophyll concentration in leaves, leaf area index, biomass, and plant productivity. We integrated both stereo vision and NDVI techniques and developed a stereo spectral imaging (SSI) system for chlorophyll and biomass quantification of plants in 3D space. We used a stereo vision camera system and custom designed a dual-wavelength filter system at 690 nm and 750 nm to develop the SSI system. Calibration techniques for NDVI computation using two spectral band images were developed by referencing a diffuse reflectance panel. We also developed a texture mapping technique for rendering NDVI values in 3D space. In this paper, the performance of the SSI system was evaluated with an artificial plant with spectral properties similar to real plant's green leaves.**

I. Introduction

Imaging systems exploiting both stereo vision and spectroscopy have been developed for various agricultural and remote sensing applications and space missions, e.g. rice seedling row detection [1], wheat plant geometric properties [2], autonomous vehicle navigation [3], digital elevation maps for land use/cover databases [4] and the Surface Stereo Imager for the Phoenix Mars Lander of NASA [5]. In this paper, a stereo spectral imaging (SSI) system collecting and processing spatial and spectral information from two camera views and two spectral bands was developed to study plant health.

3D measurements of whole plant canopies using stereo vision may provide detailed spatial structure information about plant health, disease propagation path, or phenotypes of plants used in biomass production such as cotton, poplar and switch grass, and also can complement existing X-ray systems for the below-ground part. In addition to the characterization of plants with 3D spatial information, spectral information can also biochemically characterize plants growth and health. For example, the normalized difference vegetation index (NDVI) using two spectral reflectance responses at red and near-infrared (NIR) spectral regions has been widely used in remote sensing as an index to estimate various vegetation properties including chlorophyll concentration in leaves [6], leaf area index [7], biomass and plant vigor [8], plant productivity [9], and stress [10]. We integrated both stereo vision and NDVI techniques and developed a dual-band stereo spectral imaging system for chlorophyll and biomass quantification of plants in 3D space.

Our main objective was to develop a SSI system that could monitor whole plant canopies in 3D. The specific goals were to reconstruct plant surfaces in 3D space via stereo vision, to acquire spectral features via NDVI, and to map the spectral features back onto the 3D data via computer-graphics rendering.

II. Materials and Methods

A. Overview of Stereo Spectral Imaging (SSI) System

Stereo vision is a cost-effective and rapid imaging technique for measuring and reconstructing object structures in 3D. The fundamental idea of stereo vision is based on the intersection of rays from two views in 3D, assuming that camera intrinsic parameters and the baseline of cameras (relative positions of the cameras) are known (or can be estimated). If an object feature in 3D is visible to both camera views, the feature may appear at different locations on the images captured by the cameras. The amount of spatial displacement, called disparity, is an important cue to obtain the depth of the object feature. Then, depth information is computed directly from disparity values [11]. In practice, disparity estimation is performed by a search algorithm called stereo matching to find correspondence between two separate images. Stereo matching methods to solve this correspondence problem can be categorized into two groups: region correlation-based and feature-based stereo matching. The region correlation-based stereo matching algorithm finds correspondence for every pixel in the image by using pre-determined search windows to compute correlation whereas

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_31, © Springer Science+Business Media B.V. 2010

the feature-based method finds correspondence only for a sparse set of image features (e.g. edges, lines and corners). Two common criteria for correlation computation are the sum of absolute difference (SAD) and the sum of squared difference (SSD). The SSI system was based on the region correlation-based stereo matching algorithm and the SAD criterion.

Live green plant canopies absorb solar radiation at the blue and red spectral regions for photosynthesis whereas most of the radiation at the spectral region longer than 700 nm is scattered. Hence, images of green plants appear bright in the near-infrared (NIR) spectral region and dark in the red spectral region. The NDVI was developed to utilize the properties of the different light energy absorption characteristics of plants at the NIR and red spectral regions. The NDVI at each image pixel is obtained by

$$NDVI(x,y) = \frac{\tilde{I}_{\mathrm{NIR}}(x,y) - \tilde{I}_{\mathrm{RED}}(x,y)}{\tilde{I}_{\mathrm{NIR}}(x,y) + \tilde{I}_{\mathrm{RED}}(x,y)}. \quad (1)$$

where $\tilde{I}_{\mathrm{RED}}(x,y)$ and $\tilde{I}_{\mathrm{NIR}}(x,y)$ refer to the calibrated (or normalized) relative reflectance values at the visible (red) and near-infrared (NIR) spectral regions. The NDVI value itself varies between -1.0 and +1.0. In practice, a value less than 0.2 usually means unhealthy or sparse vegetation and a value close to +1 (more than 0.6) means very healthy green leaves or very dense vegetation. NDVI values for vegetation typically range between 0.2 and 0.7. A zero or less than zero means no vegetation. However, actual wavelengths and bandwidths for NDVI calculations varied with different sensors and applications.

The SSI system consisted of a stereo vision camera (Bumblebee XB3, Point Grey Research Inc., Richmond, BC, Canada), a dual-wavelength filter system at 690 nm (red) and 750 nm (NIR) for NDVI computation, a PC and software. The Bumblebee XB3 is a 3-CCD monochrome stereo camera with the IEEE-1394b communication link (800 Mbits/s). The three lenses were horizontally placed 12 cm apart, thus three different combinations of baselines were selectable by the user via software: one wide-baseline (24 cm) and two narrow-baselines (12 cm). The lenses were sealed inside the camera unit. The focal length of each lens was set to 3.8 mm. The horizontal field of view of each lens was 70°. The camera was factory pre-calibrated for lens distortions and camera misalignments. Therefore, there was no need for in-field camera calibration and images were aligned within 0.05 pixel RMS error. The maximum image resolution per lens was 1280x960. Smaller resolutions were selectable by the user via software in order to speed up the stereo processing and to increase the robustness of the stereo matching process. The frame transfer rate was 15 frames per second at the full-resolution.

Optical band-pass filters (50.8 mm x 50.8 mm) were used for the NDVI computation. The center wavelengths of the near-infrared and red band-pass filters were 750 nm and 690 nm (Part No. 20BPF10-690 and 20BPF10-750 respectively from Newport Corporation, Irvine, CA) and two sets of these filters were used. The bandwidths (full width at half maximum) were 13 nm (NIR) and 11 nm (red). Two filter holders were designed and attached to the outside of the camera body such that the filters were placed in front of the lenses to configure the 24 cm baseline setup (Fig. 1).

Fig. 1. Stereo Spectral Imaging setup.

The selection of the NIR filter at 750 nm was to avoid the strong signal attenuation due to the oxygen absorption of natural sunlight energy at 760 nm in the consideration of the possibility of deploying the SSI system outdoors and the low quantum efficiency of the CCD sensor beyond 800 nm. Two tungsten halogen lamps (AC) were used as the lighting source in a laboratory setting for this study.

The software was written in C++ and the C++ libraries for stereo image acquisition and processing were provided in the Point Grey Research's Software Development Kit (SDK). The graphical user interface of the software was written in the Microsoft Foundation Class 6.0 (i.e. Visual C++ 6.0). An open source computer vision library OpenCV 1.0 was used to implement image processing algorithms mentioned in the paper. OpenGL [12] codes were written for implementing 3D rendering and texture mapping. Fig. 2 shows a screen capture of the software's graphical user interface.

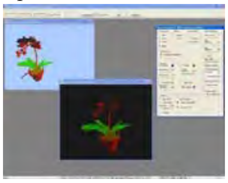


Fig. 2. Software Interface Design.

B. *Stereo Image Acquisition and Processing*

For each experiment, 3 sets of stereo images were captured: raw (i.e. no filter was used), red filters on each lens, and NIR filters on each lens (see Fig. 3). Each set was acquired one at a time by manually removing or changing to the proper filters at the filter holders as shown in Fig. 1.

(a)

(b)

(c)

Fig. 3. Raw (a), red (b) and NIR (c) stereo images.
(Note: the right camera view is located on the left of each stereo image)

The camera parameters and lighting conditions were not changed during the data acquisition period (except for a flat field calibration to be mentioned later). The captured three stereo images (raw, red and NIR) were used for the following tasks. First, the raw stereo image was used for stereo matching leading to 3D reconstruction. Second, both NIR- and red-filtered stereo images were used for NDVI computation. Third, the NIR-filtered stereo image was used for foreground object segmentation because background clutters were better suppressed by the NIR filter.

The NDVI computation required the intensity calibration (or called normalization) to obtain the relative reflectance at each pixel. We implemented a relative reflectance (RR) calibration method for the SSI system. This method was to calibrate every single pixel on the image by using a 99% diffuse reflectance panel which completely covered the plant object in the field of view. The relative reflectance calibration equation at each pixel required to collect a reference image and a dark current image, and was calculated by

$$\tilde{I}_{measurement}(x,y) = \frac{I_{measurement}(x,y) - I_{dark\ current}(x,y)}{I_{99\%\ reflectance}(x,y) - I_{dark\ current}(x,y)}. \quad (2)$$

The RR calibration method may be impractical for outdoor applications because it is not possible to put a reflectance panel large enough to cover the area in the entire field of view. Thus, we developed a flat field (FF) calibration technique more suitable for outdoor imaging of plants. The original FF calibration technique [13] was based on an assumption that the relative reflectance of a calibration target remained unchanged over wavelengths (i.e. red and NIR spectral regions). For the FF calibration, a reflectance target having very little variation in reflectance over wavelengths was placed in the scene and the digital number at each pixel was normalized to the mean value of all pixels in the region of interest (ROI) of the calibration target. The normalization was obtained by

$$\tilde{I}_{NIR}(x,y) = \frac{I_{NIR}(x,y)}{\bar{I}_{A,NIR}},\ \tilde{I}_{RED}(x,y) = \frac{I_{RED}(x,y)}{\bar{I}_{A,RED}},. \quad (3)$$

where $\bar{I}_{A,NIR}$ and $\bar{I}_{A,RED}$ were the average intensity values of the ROI (i.e., the target area A) at the NIR and red spectral bands, respectively. If $\bar{I}_{A,NIR}$ and $\bar{I}_{A,RED}$ are equal (i.e. a constant $\bar{I}_A$) and (3) is plugged into (1), then the constant $\bar{I}_A$ will be cancelled out and the un-calibrated digital numbers, $I_{RED}(x,y)$ and $I_{NIR}(x,y)$ will replace the normalized values $\tilde{I}_{RED}(x,y)$ and $\tilde{I}_{NIR}(x,y)$ in (1). For making the average intensity values of the ROI equal (or approximately equal) between the two spectral bands, we implemented an algorithm to adjust either a shutter time or a camera gain of the red-band. More specifically, we captured the NIR-band image first and then captured the red-band image by changing a camera parameter (a shutter time or a gain) such that the average intensity value of the ROI in the red-band image becomes (approximately) equal to that in the NIR-band image. The algorithm found an optimal camera parameter for the red-band capture by minimizing the sum of the absolute intensity difference between the ROIs of two bands. Then, the computation of the NDVI became possible with raw digital numbers in the place of normalized values, as defined by

$$\tilde{I}_{NIR}(x,y) = I_{NIR}(x,y),\ \tilde{I}_{RED}(x,y) = \hat{I}_{RED}(x,y),. \quad (4)$$

where $\hat{I}_{RED}(x,y)$ is the intensity value of the red-band image captured after the adjustment of the camera parameter.

Stereo processing in the SSI system was implemented by using the library functions provided in the Point Grey Research's SDK and OpenCV. The following steps were applied to all 3 stereo images:

a) After two images (8-bit monochrome 1280x960) was captured, they were interleaved into a single image according to a pre-determined image format (2560x960) in order to transfer the data according to the DCAM specification. The acquired stereo images were then pre-processed by a low-pass filter and spatially rectified to un-warp the lens distortions. Without low-pass filtering, the image rectification would produce aliasing effects. After the rectification, the radial image patterns were straightened. This process was applied to all stereo images whether the band-pass filters were used or not.

b) The rectified NIR-band stereo image was sent to an image segmentation function. For segmentation, an intensity threshold was applied to the rectified stereo image for suppressing out background clutters. Then, a morphological open filter was applied to remove small and isolated background clutters. The remaining clutters, if remained, were removed manually. Then, this segmentation image was used as a binary mask image to be applied to all stereo images.

c) The raw stereo image masked by the segmentation map was sent to the stereo matching algorithm to obtain 3D information of the object in the scene. The stereo matching was based on the sum of absolute difference (SAD) correlation method for estimating disparity at every pixel in the reference view image. The right camera view was used as the reference view image. The secondary (optional) stereo matching was based on the edge detection that utilized the relative difference in intensity values. The output of the stereo matching algorithm was a disparity image. When displaying it, each pixel on the disparity image was mapped by a look-up table representing a close object feature with red and a far-away object feature with blue. The disparity image was represented with either an 8-bit (full-pixel displacement) or 16-bit (sub-pixel displacement) precision.

d) The NIR and red images were input to the NDVI computation engine. Again, they were also masked by the segmentation map. A pseudo-color image was generated by a color look-up table in order to display the resulting NDVI image. The color look-up table of the NDVI values was made according to the following rule:

- If NDVI is less than or equal to 0.2, then grey value. The black refers to -1.

- If NDVI is greater than 0.2 and less than or equal to 0.8, then mixture of pure red (0.2) and pure green (0.8).
- If NDVI is greater than 0.8, then mixture of green and pure blue (1.0).

The floating-point NDVI values were also saved for future numerical quantification.

C. Rendering of NDVI Image in 3D

After stereo processing, each pixel in the right camera view (main view – 2D) was mapped into a point in the camera coordinate system (3D) where its origin was at the focal point of the right camera (please note that by convention this image was located on the left of each stereo image as shown in Fig. 3). The set of these points in 3D were called point clouds. Obtaining point clouds from stereo matching was the first step for 3D surface reconstruction. Simply, the point clouds were nothing but a set of 3D coordinates representing surface points in 3D. Thus, it was important to spatially connect these surface points in order to create an approximate surface representation. This representation was obtained by a geometric triangular mesh model. Rendering of point clouds and triangular meshes in 3D was implemented by OpenGL functions. The quality of the 3D mesh representation was greatly affected by the quality of generated cloud points. Typically, point clouds needed to be cleaned and edited to remove duplicated and unreferenced points. For this task, when necessary, we removed isolated and tiny surface features by applying a median filter directly to a disparity image. The 3D coordinates of the point clouds were also saved as a standard file format (xyz format) so as to be used in the free open-source 3D data processing software (MeshLab) for off-line processing of these point clouds.

Image rendering based on OpenGL was implemented for mapping the computed NDVI values onto 3D point clouds as well as a mesh model in the camera coordinate system such that a 3D virtual image was generated at any given viewpoint. Also, 3D anaglyph NDVI images were generated in real-time to provide a 3D stereoscopic perception when viewed with two color (red and green) glasses. Fig. 4 is the data flowchart of the 3D rendering algorithm of the SSI system.

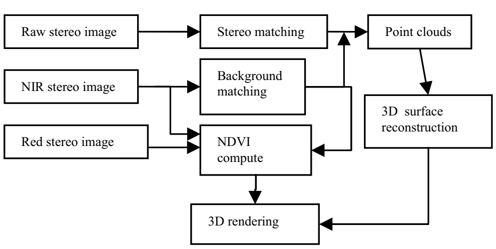


Fig. 4. Data flow from stereo image acquisition to 3D rendering.

D. Other materials needed for experimentation

A 99% diffuse reflectance target (SRT-99-180, Spectralon, 18”x18”, Labsphere, North Sutton, NH) was used as the reference material for the relative reflectance (RR) calibration (Eq. 2). Another 99% diffuse reflectance target (SRT-99-050, Spectralon, 5”x5”, Labsphere) was used as the reference material for the flat field (FF) calibration (Eq. 4). To evaluate the RR calibration technique, we used seven target materials: a 5”x5” multi-step diffuse reflectance target (SRT-MS-050, Labsphere) having four sub-panels of 99%, 50%, 25% and 12% nominal reflectance values, two vinyl sheets used for military camouflages (uniform dark and light green) and their back sides (black). The nominal reflectance values of the vinyl sheets were measured by a four channel fiber-optic spectrometer (SQ2000, Ocean Optics, Dunedin, FL). For evaluating the accuracy of the computed NDVI values, we used the three camouflage vinyl sheets.

An artificial orchid plant (Fig. 1) was used for the SSI system development because its leaves showed similar NIR spectral reflectance as compared to real leaves. Also it was essential to have a reference plant during the system development cycles because the artificial plant was not perishable and provided good references over time.

III. Results

A. Calibration and NDVI Computation

The performance of the RR calibration technique was evaluated by using seven test materials placed in the camera field of view: the multi-step reflectance target (99%, 50%, 25% and 12%) and the two vinyl sheets (dark and light green) and their back side (black). The performance of the RR calibration was compared to spectrometer measurements. The calibration of the spectrometer measurements was done on the fly by the Ocean Optics’ software using its own dark-current & reference (Spectralon) spectra. Calibration values at 689.92 nm and 749.99 nm of the spectrometer were compared against the calibration values of the SSI system at the NIR and red bands. Fig. 5 shows the scatter plot of calibrated image and spectrometer measurements of the test materials. Each point in the figure represented an average calibration value of one particular material at one of the two spectral wavelengths, thus 14 points were obtained. A regression line was overlaid to represent the relationship between the two instruments. The R^2 value (goodness-of-fit) of the linear regression was 0.9912. Overall, the measured values from the imaging system were slightly lower than the spectrometer measurements. But, the difference was consistently small across the entire calibration value range from 0 to 1.

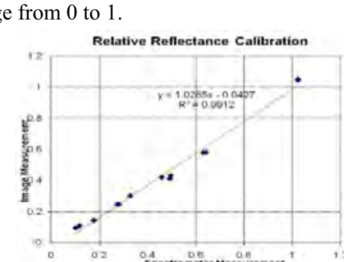


Fig. 5. Scatter plot of calibrated relative reflectance measurements from the SSI system and the spectrometer. The coordinates represented normalized relative reflectance values. The linear regression line was drawn.

The calculated NDVI values were evaluated in terms of the RR calibration and the FF calibration by comparing them with NDVI values calculated by the spectrometer. The vinyl sheets (dark green, light green, black) were used as test materials. Fig. 6 shows the scatter plot of the NDVI values computed by the SSI system and the spectrometer. Numerical numbers are presented in Table I. A linear regression model was applied to find the linear relationship between two instruments in NDVI computation. The R^2 values of the linear regression were 0.9974 (RR), 0.9921 (FF-Shutter), and 0.9919 (FF-Gain). The performance of the RR calibration was better than the FF calibration methods. However, the FF calibration technique also produced NDVI values close to what the spectrometer produced. In fact, as shown in Fig. 6, the performance gap among three calibration methods decreased as the NDVI value increased. The overall performance of the SSI system adopting either of three different calibration techniques was similar to that of the spectrometer.

The NDVI values of the sample plant were also measured by the spectrometer. Table II summarizes the NDVI values measured at various points of the plant. Five different points were measured for leaf, two different points were measured for stem, and three different points were measured for flower. As shown in Table II, the average and the standard deviation values were obtained. The average NDVI value of the leaf was 0.619 which was similar to normal healthy leaves of green plants. The average NDVI value of the stem was 0.1 less than 0.2. The average NDVI value of the flower (0.041) was even lower than the stem's.

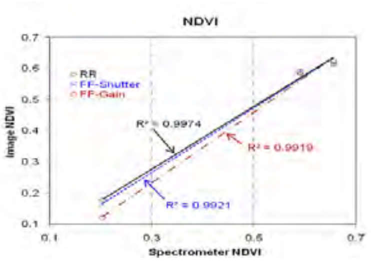


Fig. 6. Scatter plot of NDVI values of the SSI system and the spectrometer. The linear regression line was drawn. The "RR" and "FF" refer to the RR and FF calibration-based NDVI calculations, respectively. The "Shutter" and "Gain" refer to one of the FF calibration techniques: shutter time control and gain control.

Table I.
Average NDVI values of the vinyl sheets: SSI system and spectrometer measurements

	Spectrometer-based NDVI	RR-based NDVI	FF-Shutter-based NDVI	FF-Gain-based NDVI
Dark Green	0.656	0.623	0.614	0.612
Light Green	0.591	0.583	0.591	0.588
Back-side	0.203	0.177	0.163	0.121

Table II.
NDVI values of the test sample: spectrometer measurements

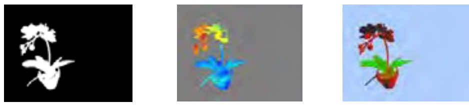
	Reflectance at 689.92 nm	Reflectance at 749.99 nm	NDVI
Leaf	0.099 ± 0.063	0.372 ± 0.110	0.619 ± 0.136
Stem	0.135 ± 0.065	0.160 ± 0.048	0.100 ± 0.096
Flower	0.244 ± 0.091	0.261 ± 0.078	0.041 ± 0.059

B. *3D Surface Reconstruction and Rendering*

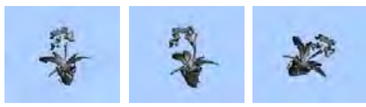
The parameters used for stereo processing of the sample plant were: stereo search window size, 11x11; maximum disparity search range, 302; minimum disparity search limit, 185; image resolution, 512x384; intensity threshold for object segmentation, 8; morphological open filter size, 3. The 24 cm baseline was used for stereo image acquisition and processing. The synthetic diffuse white light was used for 3D rendering. In OpenGL, the orthographic projection transformation was adopted for 3D rendering. The surface normal vector was computed for providing OpenGL with the correct direction of each mesh's surface. The surface normal vector was positive when the vector direction was coming off the surface (not getting into the surface). The 2D NDVI image (pseudo-color) was input to the OpenGL function for texture mapping of the 3D meshes. The rotation, translation, and zoom in/out by a computer mouse was implemented in the software so that the user can freely move around the 3D object.

Fig. 3 showed the acquired stereo images: (a) Raw, (b) Red and (c) NIR, where we can see that the NIR image suppressed the background clutters shown in the Raw image. The NIR image also showed good contrast for segmentation of the plant on the foreground. Fig. 7(a) shows the segmentation result for the plant. This segmentation mask was applied to the subsequent disparity and NDVI images, as shown in Figs. 7(b) and 7(c), respectively. The disparity map was a pseudo color image similar to temperature. The hotter (red) pixels meant closer to the camera (or viewer). The cooler (blue) pixels meant away from the camera. The NDVI image shown in Fig. 7(c) was just a 2D image. As mentioned in Table II, the average NDVI value for leaves was 0.619 which should be close to green in the NDVI-mapped color image. The blue in Fig. 7(c) meant NDVI values were greater than 0.8. The stems painted with the red color meant NDVI values close to 0.2. The gray color shown in flower areas meant NDVI values below 0.2. The NDVI values at different parts of the plant were not quantified in this paper but it can be possible by an image segmentation technique to partition the plant into multiple segments like leaves, stems and flowers. The 3D rendering results are represented as point clouds in Fig. 7(d), as triangular meshes in Fig. 7(e) and as texture maps of NDVI values in Fig. 7(f). The overall quality of the 3D rendering was acceptable with this sample. The imaging system, however, showed difficulties with highly textured plants, whether real plants or not, and still needs to make some improvement.

The quality of the 3D rendering was greatly dependent on the performance of the stereo matching algorithm. The current stereo vision camera system showed a difficulty in capturing small surface features from highly textured areas due to the inherent limitation of the stereo matching algorithm. Thus, the use of the camera on small plants was challenging. The most important parameter for depth estimation was the min-max range of the disparity values to search for. In order to improve the quality of the 3D rendering, the accuracy of the depth estimation needs to be improved by a better stereo matching algorithm or the point clouds may be cleaned by multiple stereo images taken at different viewpoints.

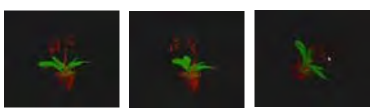


(a) Segmentation mask (b) Disparity map (c) NDVI image



(d) Point cloud rendering at different viewpoints

(e) Triangular mesh rendering at different viewpoints



(f) NDVI texture rendering at different viewpoints

Fig. 7. Screen captures of the processed images and 3D rendering results. The rendering results at different viewpoints were captured.

IV. Conclusions

We developed a stereo spectral imaging (SSI) system for quantifying plants' health by rendering normalized difference vegetation index (NDVI) values in 3D. We used a stereo vision camera system and custom designed a dual-wavelength filter system at 690 nm and 750 nm to develop the SSI system. The bandpass stereo images of the SSI system were calibrated for the NDVI computation. A new calibration technique based on the flat field calibration was developed for outdoor imaging. The NDVI values measured by the SSI system were compared with the NDVI values measured by a spectrometer. Both performed very similarly (the best goodness-of-fit, R2 was 0.9974). We also developed a texture mapping technique for rendering NDVI values in 3D. The performance of the SSI system was evaluated with an artificial plant showing spectral properties similar to green leaves of real plants. The test of the SSI system with real plants is planned.

References

[1] Y. Kaizu and K. Imou, "A dual-spectral camera system for paddy rice seedling row detection," *Computers and Electronics in Agriculture,* vol. 63, no. 1, pp. 49-58, 2008.

[2] H. J. Andersen, L. Reng and K. Kirk, "Geometric plant properties by relaxed stereo vision using simulated annealing," *Computers and Electronics in Agriculture,* vol. 49, no. 2, pp. 219-232, 2005.

[3] M. Kise and Q. Zhang, "Creating a panoramic field image using multi-spectral stereovision system," *Computers and Electronics in Agriculture,* vol. 60, no. 1, pp. 67-75, 2008.

[4] F. Paul, C. Huggel, and A. Kääb, "Combining satellite multispectral image data and a digital elevation model for mapping debris-covered glaciers", *Remote Sensing of Environment,* vol. 89, no. 4, pp. 510-518, 2004.

[5] M.T. Lemmon, P.H. Smith, C. Shinohara, R. Tanner, P. Woida, A. Shaw, J. Hughes, R. Reynolds, R. Woida, J. Penegor, C. Oquest, S.F. Hviid, M.B. Madsen, M. Olsen, K. Leer, L. Drube, R.V. Morris, D. Britt, "The Phoenix surface stereo imager (SSI) investigation," in *Proc. 39th Lunar and Planetary Science Conf. (Lunar and Planetary Science XXXIX)*, vol.1391, pp. 2156-2157, 2008.

[6] C.L. Jones, N.O. Maness, M.L. Stone, and R. Jayasekara, "Chlorophyll estimation using multispectral reflectance and height sensing", *Transactions of the ASABE,* vol. 50, no. 5, pp. 1867-1872, 2007.

[7] R. Drissi, J.-P. Goutouly, D. Forget, and J.-P. Gaudillere, "Nondestructive measurement of grapevine leaf area by ground normalized difference vegetation index," *Agronomy Journal,* vol. 101, pp. 226-231, 2009.

[8] R.B. Boone, K.A. Galvin, N.M. Smith, and S.J. Lynn, "Generalizing El Nino effects upon Maasai livestock using hierarchical clusters of vegetation patterns," *Photogrammetric Engineering and Remote Sensing,* vol. 66, no. 6, pp. 737-744, 2000.

[9] J.P. Wang, P.M. Rich, K.P. Price, and W.D. Kettle, "Relations between NDVI and tree productivity in the central Great Plains," *International Journal of Remote Sensing,* vol. 25, no. 16, pp. 3127-3138, 2004.

[10] R.E. Plant, D.S. Munk, B.R. Roberts, R.L. Vargas, D.W. Rains, R.L. Travis, and R.B. Hutmacher, "Relationships between remotely sensed reflectance data and cotton growth and yield," *Transactions of the ASAE,* vol. 43, no. 3, pp. 535-546, 2000.

[11] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision,* Upper Saddle River, N.J.: Prentice Hall, 1998.

[12] D. Shreiner, M. Woo, J. Neider, and T. Davis, *OpenGL(R) Programming Guide: The Official Guide to Learning OpenGL(R), Version 2.* 5^{th} ed., Reading, MA: Addison-Wesley Professional, 2005.

[13] D.A. Roberts, Y. Yamaguchi, and R.J.P. Lyon, "Comparison of various techniques for calibration of AIS data," in *Proc. of the Second Airborne Imaging Spectrometer Data Analysis Workshop,* JPL Publ. 86-35, pp. 21-30, 1986.

Transition Network for Automation Modeling: A Case Study

Yuval Cohen*

Department of Industrial Engineering
The Open University of Israel
Raanana, Israel, 43107

Ming-En Wang, Bopaya Bidanda

Department of Industrial Engineering
The University of Pittsburgh
Pittsburgh, USA, 15261

Abstract *–This paper illustrate through a case study the advantages of the proposed Transition Network (TN) for modelling discrete automation on the shop floor. TN is associated with high-level state diagram The proposed technique can easily simulate and emulate the automation logic and can be automatically translated to PLC (Programmable Logic Controller) code.*

Keywords: Transition Network, PLC, Automation, Ladder-diagram

I. Introduction

The traditional modelling approach for state diagrams of discrete automation approach typically characterize a state as any possible combination of inputs and outputs [1,2,3]. The emphasis is on the states. Each change in this combinatory design is a transition of one or more inputs and/or outputs [2, 3,4]. Typically the logic of the generated models grows rapidly with the number of inputs and outputs. Moreover, tracking the logic is a very difficult task [2,4,5]. In this paper we suggest a different approach with intuitive representation of cause and effect which facilitate tracing of logic. In our model the states represent only the output combinations. These are either machine actions or some machine states of rest. The transitions are invoked by input changes and result in changes to some outputs.

The technique presented in this paper not only relieves the automation engineer from dealing with Programmable Logic Controller (PLC) code when constructing the logic, but also in later stages, when changes are required. This approach offers intuitive action representation facilitating verification and validation. We also adopt the convention that there can be only one task (state) node following each transition, and that all the required resources for each task (state) are available by incorporating the inputs related to their availability into the transition conditions.

II. The Example Case Study

For the case study a drill-press system is chosen to illustrate the remaining steps of the proposed technique. The system is adapted from [6]. Figure 1 depicts the main components and the layout of the automated drill-press that will be used throughout this paper. The drill press consists of an electric drill mounted on a movable platform. The platform is attached to a piston of a hydraulic cylinder. The movable platform moves vertically between the up position and the down position. The limit switch LS1 senses when the drill is in the up position and the limit switch LS2 senses when it is in the down position. The hydraulic cylinder is controlled by a hydraulic three-position, four-way valve with two solenoids. Solenoid A drives the hydraulic cylinder down. Solenoid B drives it up. When both solenoids (A and B) are deenergized the platform is stationary.

A stationary table supports the workpiece and a pneumatic clamp clamps the workpiece against a bracket. Limit switch LS4 is actuated when the clamp is in the clamped position. Limit switch LS3 is actuated when the clamp is in the unclamped position. An ejector is also attached to the table (oriented 90 degrees from the clamping cylinder). The ejector ejects the workpiece when the drilling is finished. Limit switch LS5 is actuated when the ejector is retracted and LS6 is actuated when the ejector is in the eject position.

* Corresponding author: Full name: Yuval Cohen, Academic degree: PhD, Working place: The Open University of Israel, Research field: Automation and production modeling, Post address: 108 Ravutski, P.O.Box 808, Raanana, Postcode 43107, Tel.: (972) 9-778-1883; Fax: (972) 9-778-0668; E-mail: yuvalco@openu.ac.il

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_32,

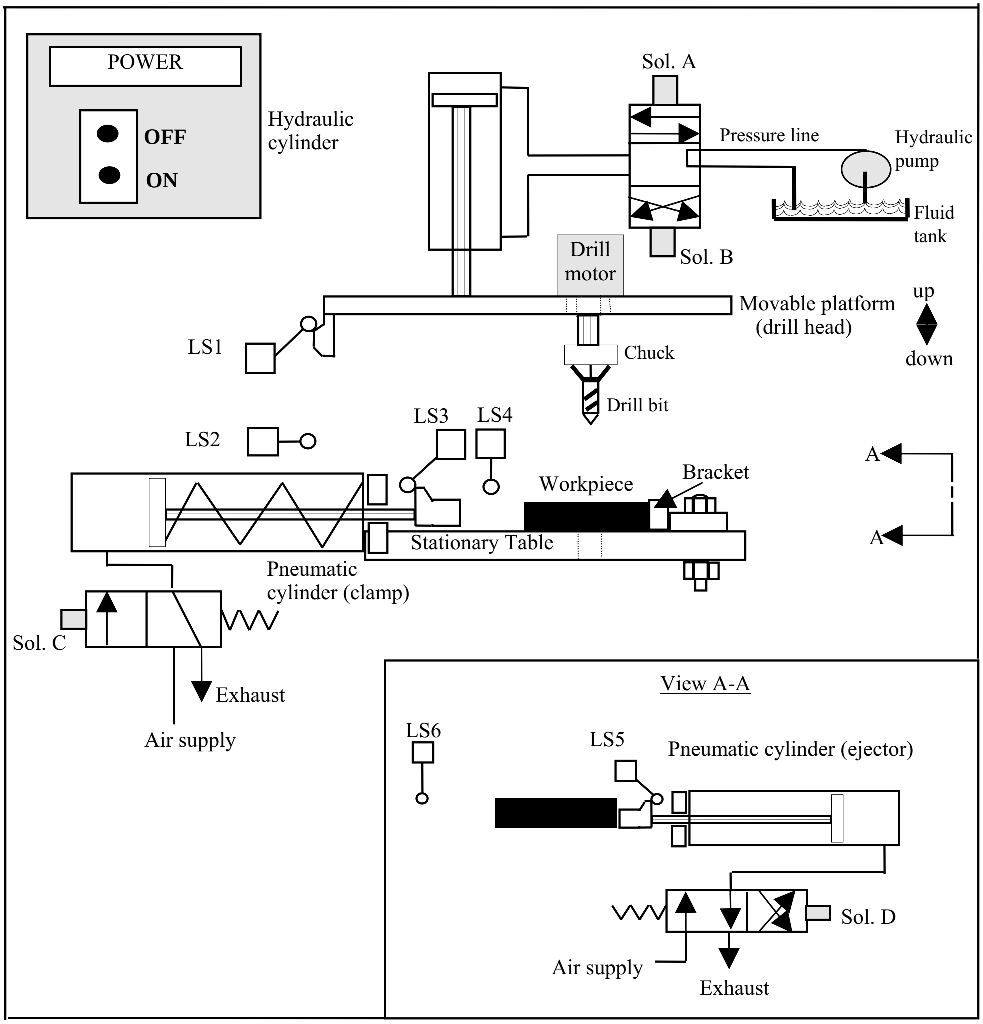


Fig. 1. The case study system: an automated drilling machine (adapted from [6])

Next, the operation of the system described in figure 1 has to be described by a state diagram that is constructed as explained in section II. Note that the states in section II are related to the status of outputs only.

II. State-Diagram Construction

The aim of the state diagram is to draw a framework for the transition network. This is very different from a regular state diagram. The states represent actions or no action of the machines and are therefore related only to combinations of the machines outputs. Ignoring inputs dramatically simplifies the state diagram. In a later stage only the necessary input combinations are considered. The design of actions (legal states) is facilitated by analysing the steps.

For the case study of figure 1, the eight legal system states for the drill press example are as follows:

1) **OFF**
2) **Idle** (only the hydraulic cylinder is ON)
3) **Ready** - (similar to Idle but the drill motor is ON and the clamp is in the clamped position)
4) **Descend** - (similar to Ready but Solenoid A is activated)
5) **Ascend** - (similar to Ready but Solenoid B is activated can be with or without clamping)
6) **Release clamp** - (similar to Ascend but Solenoid C is deactivated to release the clamp)
7) **Eject workpiece -** (similar to Release but Solenoid D (the ejector) is activated)
8) **Retract ejector** - (similar to Release but Solenoid D is deactivated)

At this point a state diagram can be constructed. Figure 2 depicts the state diagram for the drill press specification. States are denoted by numbers and transitions by capital letters.

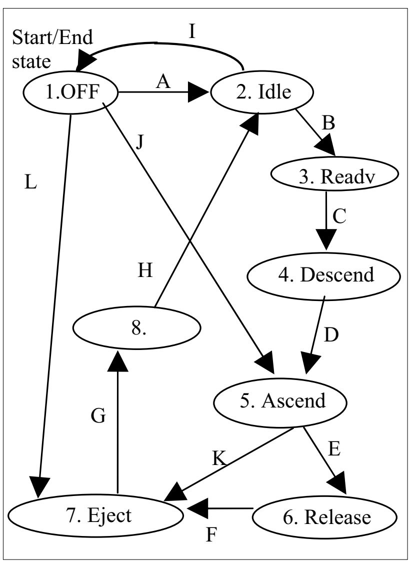


Fig. 2. The case study State Diagram

The states do not include all the details regarding inputs, outputs, and variables. These details are embedded in the TN. Next, essential data manipulation is described to enable the generation of the TN. The main input required for the proposed scheme, is a specification of the output and timers status for each state. An example for such input (for the drill-press machine) is depicted in Table 1. This specification depends heavily on the specific system at hand, and must be done carefully and correctly.

TABLE I
STATUS OF OUTPUTS FOR STATES IN FIG. 2.
(Entries representing actions or active timers are shaded.)

	OUTPUTS							
	Solenoids							
State	A	B	C	D	Pump motor	Drill motor	Coo -lant	Tim- er 1
1	off	off	off	off	off	off	off	off
2	off	off	off	off	on	off	off	On
3	off	off	on	off	on	on	off	off
4	on	off	on	off	on	on	on	off
5	off	on	on	off	on	on	off	off
6	off	off	off	off	on	off	off	off
7	off	off	off	on	on	off	off	Off
8	off	off	off	off	on	off	off	Off

Each transition in the state diagram has origin state and destination state. The status changes between the origin and destination outputs and timers are the transition effects. Therefore, by comparing the lines in Table 1 that relate to the origin and destination states, the effects could be generated automatically.

Transition effects are summarized in Table 2. For example, it is easy to see that the changes between lines 1 and 2 in table 1 are in the pump motor and timer. These changes appear in line one of Table 2.

TABLE II
THE EFFECTS OF TRANSITIONS OF FIG. 2.

Transition	**Its effect/s**		
A (1→2)	Pump motor ↑	Timer ↑	
B (2→3)	Drill motor ↑	Solenoid C ↑	
C (3→4)	Solenoid A ↑	Coolant ↑	
D (4→5)	Solenoid A ↓	Coolant ↓	Solenoid B ↑
E (5→6)	Solenoid B ↓	Solenoid C ↓	Drill motor ↓
F (6→7)	Solenoid D ↑		
G (7→8)	Solenoid D ↓		
H (8→2)	Timer ↑	Reset ↓	Done ↑
I (2→1)	Pump motor ↓	PROCESS2↓	Done ↓
J (1→5)	Pump motor ↑ Reset↑	Solenoid B ↑	Drill motor↑
K (5→7)	Solenoid B ↓	Solenoid D ↑	Drill motor ↓
L (1→7)	Pump motor ↑	Solenoid D ↑	Reset ↑

The generation of significant portions of Tables 1 and 2 can be automated. In addition to the elements discussed so far, there are external variables originating in higher control levels. Moreover, different history of past events may set different modes of operation. Such modes are identified by internal variables. The additional variables for the drill-press example are as follows:

External Variables

1. PROCESS2 signifies that the resources for task 2 are occupied. It is turned ON when the workpiece is ready to be drilled (the robot finished placing the workpiece at the drill-press). It shuts OFF when the drill is ready to execute a new task.
2. FINISH2 signifies that the drilling task is finished. It is turned ON when the drilling task is finished, and turned OFF after invoking the next transition.

Internal Variables

1. **Timer1** is used to warm up the hydraulic engine and cool it down (turned ON by transitions A and H).
2. **Reset** variable (signifies that the system is recovering from power failure). Reset is turned ON whenever power is turned ON (transitions J and L) and turned OFF after workpiece ejection (transition H).
3. **Done** variable signifies that the next time the system will be idle is a cool down period followed by system shut OFF.
4. Eight state-variables: **ST1,ST2, ST3, ST4, ST5, ST6,ST7, ST8,** signifying that the system is at a certain state.

III. Transitions Network Generation

The state diagram is further exploded into a new type of a graphical scheme named Transition Network (TN). TNs describe the changes in low level elements such as inputs, outputs, and registers, required for executing the state diagram. TNs arrange the elements in a meaningful way that enables immediate comprehension of a low level code. The TNs operate much like Petri networks. However, instead of tokens, the transition is triggered by change of inputs and its effects are modeled as turning On and OFF certain outputs. The transitions are composed of the following elements: 1) places, 2) triggers, and 3) arcs. These elements are all depicted in figure 3. Each transition is activated by one or more triggers. The triggers are denoted by triangles pointing at a thick vertical line that symbolizes the transition. Places (denoted by circles) represent the inputs, outputs, events, and variables.

Events are assigned places with additional symbol to denotes the type of event (turn ON, and shut OFF). Places that use non-binary data (e.g., timers and counters) are denoted by rectangles. Additionally, places are added for logically denoting the states of the system. For example, in table III transition C from state 3 to state 4 is triggered by the corresponding ST3 and additional input change. Its effects include activating ST4 while deactivating ST3.

Two arc types used to activate triggers are as follows:
1. An enable arc (········▶) the triggers can fire only while the source place is ON.
2. A disable arc (··········●) the triggers can fire only while the source place is OFF.

Enable and disable arcs are drawn with dashed lines to denote that they do not activate or deactivate elements. A tokens is used to denote the current activated state. Two types of arcs used to identify the effects of a transition as follows:
1. Activate arc (········▶) turns ON the place when the corresponding E-Transition is activated.
2. Deactivate arc (——⊠) turns OFF the place when the corresponding E-Transition is activated.

TABLE III
CASE STUDY TRIGGER EVENTS

Trigger event	Transition	Source state: Var. (name)	Detected Inputs Change	Supplementary Identifiers
1. Start	A	ST1 (closed)	PROCESS2 $\uparrow$	
2. Reset ascend	J	ST1 (closed)	Reset$\uparrow$	LS1/**OFF**
3. Reset ejection	L	ST1 (closed)	Reset$\uparrow$	LS1/**ON**
4. Warm-up completed	B	ST2 (Idle)	Timer $\downarrow$	*Done*/**OFF***
6. Cool Down completed	I	ST2. (Idle)	Timer $\downarrow$	*Done*/**ON**** (Internal variable turned ON by Transition H)
7. Clamp	C	ST3. (Ready)	LS4$\uparrow$	
8. Down	D	ST4. (Descend)	LS2$\uparrow$	
9. Up	E	ST5. (Ascend)	LS1$\uparrow$	Reset/**OFF**
10. Reset	K	ST5. (Ascend)	LS1$\uparrow$	Reset/**ON**
11. Released	F	ST6. (Release)	LS3$\uparrow$	
12. Ejected	G	ST7. (Eject)	LS6$\uparrow$	
13. Retract	H	ST8. (Retract)	LS5$\uparrow$	

*Internal variable turned OFF by Transition I
****** Internal variable turned ON by Transition H

Each trigger is invoked by places linked to the trigger by enable or disable arcs. Note the usage of the source state (STi) variable of the transition to facilitate trigger's identification as one of the trigger's conditions. After the trigger is activated, a transition from the source state (i) to another state (j) occurs immediately. Each transition also resets the source state variable (STi) and sets the destination state variable (STj). Note that each trigger has only one transition, but a transition may have more than one trigger. Since the system starts at an initial state and the current state is updated as integral part of each transition, tracking the current state is done automatically. Moreover, the trigger identification is facilitated as a result. Trigger identification for the case study is elaborated in Table III, and is depicted in figure 3.

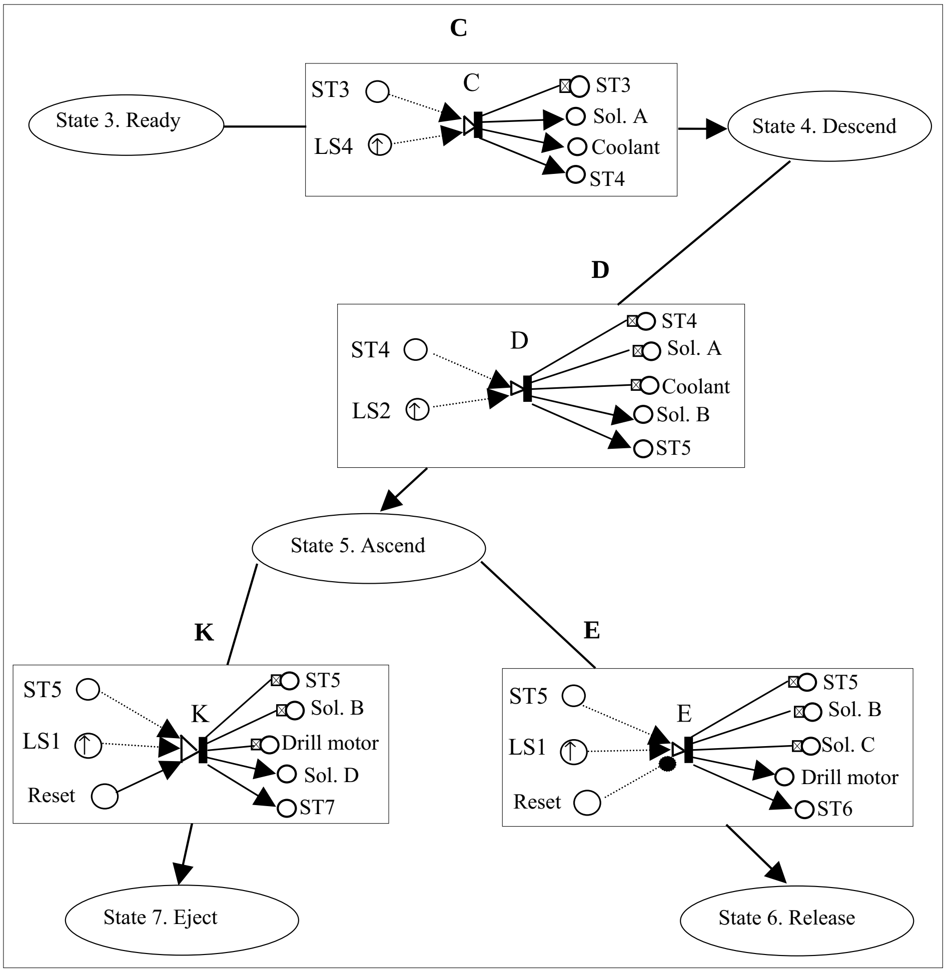


Fig. 3. A segment of Y net integrating EASD with the embedded E transitions

IV. Trnslating the TN to PLC Code

A Ladder Diagram (LD) is a popular PLC code [3,4,5,6] which was chosen to illustrate the translation of the model to PLC code. Today, most of the control programs are developed using LD [7]. The generated LD rungs are arranged in three main blocks as follows: 1) events identification 2) transition triggers, and 3) transition effects. The construction of the above three blocks is presented next.

A. Events Identification

Inputs and outputs change their voltage level when turned ON or OFF. These changes are referred as rising or falling edges. The international standard IEC 61131-3 [8] defines special LD contacts for detecting rising and falling edges. A rising edge corresponds to a place with "↓" and a falling edge to a place with "↑".

B. Transition Triggers

Each trigger activates one E-Transition. Each transition is assigned an internal variable in the LD. When the E-Transition is enabled that variable will be turned ON. In order to implement this logic, a set of rules is described as follows:

I. Each trigger forms an LD rung.

II. Each place (in E-Transition) that is input to a trigger forms a contact: (enable arc forms a normally open (NO) contact, and disable arc a normally closed (NC) contact.

III. The LD rung output is a variable that corresponds to the activated transition.

Figure 3 depicts a ladder diagram segment corresponding to the triggers of transition D. These variables are used in figure 4.

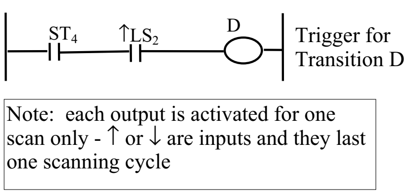


Fig. 4. Ladder Diagram segment for triggering transition D

C. Transition Effects

The rules for establishing the ladder diagram portion of transition's effects is as follows:

1. Dedicate a rung for each output place of the transition and add to it a corresponding LD output.
2. In each rung add a contact that corresponds to the relevant transition.
3. Activation arcs are translated into latched outputs, and Turn-off arcs are translated into Unlatched outputs.

V. Conclusion

This paper presents a new modeling technique for discrete control modeling technique that is intuitive (based on causes and effects) and has a potential to be used also for simulation, validation, code generation and maintenance. The technique is illustrated using a case study and shows a great potential. Such simplification could reduce the cost of discrete automation and could contribute to progress towards more flexible automation. The proposed technique could be easily translated to code of existing control equipment (PLCs). For each processing task, a simple state diagram that describes only the elementary actions of the system has to be built. From that point the process of the model generation could be automated, provided that the following input is specified: the status of outputs for each state; the related inputs of each state, and supplementary trigger identifiers.

Future research includes implementation of the proposed technique on an industrial shop floor. This would enable simulation and visualization of operation; it would assist in real-time tracking and failure analysis of the control system. Currently the construction of a translator (model to LD) is under way, to enable such future implementation.

References

[1] Ferdinand Wagner (Author), Ruedi Schmuki (Author), Thomas Wagner (Author), Peter Wolstenholme (Author) "Modeling Software with Finite State Machines: A Practical Approach", AUERBACH, 2006

[2] Peng S. and Zhou M. C., "Sensor-based Petri net modeling for PLC stage programming of discrete-event control design", Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Volume 2, 1907 - 1912, 2002.

[3] Manesis S. and Akantziotis K., "Automated synthesis of Ladder automation circuits based on state-diagrams", Advances in Engineering Software, Volume 36 , Issue 4, 225 – 233, 2005.

[4] Collins K., "PLC Programming for Industrial Automation", Exposure Publishing, 2007

[5] Bolton W., "Programmable Logic Controllers", Fourth Edition, Newnes, 2006.

[6] Bateson N. R., "Introduction to Control System Technology" (7th Edition), Prentice Hall; 2001.

[7] Lee G. B., Zandong H., Lee J. S."Automatic generation of ladder diagram with control Petri Net", Journal of Intelligent Manufacturing, 15, 2, 245-252, 2004.

[8] Karl-Heinz J. and Tiegelkamp, M., IEC 61131-3: Programming Industrial Automation System, w. CD-ROM : Concepts and Programming Languages, Requirements for Programming Systems,6^{th} *edition, Springer, 2001.*

Computational Modeling and Simulation of a New Energy-Saving Sealing Unit in Industrial Applications

Zheng (Jeremy) Li

University of Bridgeport, Bridgeport, CT 06604, USA

Abstract

This new type of magnetic energy-saving sealing system can reduce the leakage of lubrication oil pollution and media gases in general industrial machine applications including reciprocating machines, air compressors and refrigerating units. This article analyses the feasible and reliable function of this new magnetic energy-saving sealing system through computer aided modeling / simulation and prototype testing, and the results verified the reliable performance of this sealing system. The main features of this new sealing system include: better sealing performance to prevent the industrial machines from lubrication oil and gaseous leakage, energy-saving due to higher sealing functioning efficiency with less frictional work loss and reduced internal media leakage, cheaper in industrial production and manufacturing because of lower precision requirement to the surfaces of reciprocating components including pistons and shafts, more durable sealing function and longer sealing lifetime.

Introduction

The lubrication oil and gaseous leakage in industrial machines are common problems that were not resolved efficiently ([1]), and the machinery performance were therefore degraded ([2] and [3]). The research and development of new effective sealing system are being conducted in these years to improve the sealing performance ([4] and [5]).

In this new magnetic energy-saving sealing system, a rare-earth magnet is used as a permanent magnet to control the sealing capacity to solve the leakage problems in industrial machine applications. This research is conducted based on theoretic analysis, computational modeling and simulation, and prototype testing. The recent development of rare earth permanent magnets have brought the renovation and provided the magnetic products with stronger magnetic properties than that of conventional ferrite magnets. This leads to the development of high-intensity magnetic circuits that operated energy free and surpasses the electromagnets in strength and effectiveness.

The results of prototype testing indicated that this new magnetic sealing system can significantly reduce the leakage problem in industrial machine application and it can replace the regular sealing units including rubber seal, diaphragm seal, corrugated pipe seal and magnetic fluid seal.

Magnetic Circuit Analysis in Sealing System

The rare-earth magnet has the higher density of magnetic flux Br, strong magnetic field H_g, and larger product of magnetism and energy $(BH)_{max}$, and it can be used as the permanent magnet in the magnetic circuit shown in Fig. 1. All the above good features lead magnetic particles to be firmly bonded to the inside wall of magnet. This magnetic circuit shows higher Br in circuit working gap, longer and durable lifetime in sealing capacity, compact in system structure, light in unit weight, higher in performance efficiency, and stable in functioning.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_33, © Springer Science+Business Media B.V. 2010

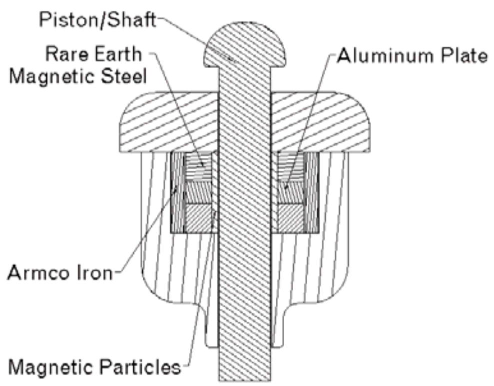


Fig. 1 Magnetic Sealing System

This new magnetic sealing system can be briefly described as follows. When piston/shaft reciprocates inside of the cylinder, lubricating oil is sealed by magnetic particles that firmly bonded to the inside wall of magnet, as oil particles move to the seal. Finally the oil droplets will drop to the main shaft chamber at the bottom of machine by its gravity and this can prevent the oil in crank chamber from entering the gaseous cylinder. Also the gaseous leakage can be prevented because the gas could not pass through the strong bonded magnetic particle layers. Two critical factors in this new magnetic sealing system design are density of magnetic flux and magnetic stability of the magnet. The magnetic flux of magnetic circuit in this sealing system should be kept steady for a long period and magnetic field of this magnet must be stable enough to resist the fluctuation caused by external/disturbed magnetic fields, temperature change, mechanical vibration/shock, and severe environment. The surplus density of magnetic flux Br, surplus intensity of magnetic field H_g, and maximum product of magnetism and energy $(BH)_{max}$ are required to keep at their peak values in this magnetic sealing system design.

The magnetic circuit of this sealing system is in static condition which can be analyzed using ampere enclosed circuit and H-B curve of this rare-earth magnet. This magnetic circuit can be considered as a series magnetic circuits mainly composed of rare-earth magnet and working gap. From Fig.1, we can get:

$$H * L + H_g * L_g = 0 \tag{1}$$

$$H * L = - \frac{Lg * \Phi}{U_0 * Ag} \tag{2}$$

Here,

H = intensity of magnetic field of magnet steel

L = length of magnet steel

H_g = intensity of magnetic field in working gap

L_g = length of working gap

Φ = magnetic flux

U_0 = magnetic conductivity of vacuum

A_g = cross section area of working gap

Assume Fm(Φ) = H * L, the intersection of Fm(Φ) and straight line – $[L_g / (U_0 * A_g)] * \Phi$ at ordinate in Fig. 2 is the magnetic flux in working gap required to be found. This gap decreases from Lg to Lg' after magnetic particles being added into the space in magnetic circuit gap. As the thickness of magnetic particles in the space between walls of magnet and cylinder changes from 0 to b, the working point of magnet moves along the straight line QK and corresponding solution of magnetic flux in working gap can be found from line QK. The computational modeling simulation verifies that the magnetic field is well maintained in this sealing system.

The coefficient of magnetic efficiency f is used to judge if the magnetic field in this sealing mechanism is properly designed.

$$f = \frac{Bg^2 * Vg}{(B * H)_{\max} * V} \tag{3}$$

Here,

B_g = density of magnetic flux in working gap

V = volume of magnet steel

V_g = volume of working gap

B = half length of working gap

The higher f value indicates more feasible and reasonable design to the magnetic circuit. The f value is normally 40% in standard conditions and the computational modeling solution shows that f value in this sealing magnetic system is 49.3% which verifies the proper magnetic circuit design in this sealing system.

Computational Modeling and Simulations

The results of computational modeling and simulation of this prototype are shown in Table 1 and 2.

Table 1. Air leakage at different piston linear speed.

Piston Linear Speed (Ft/Min)	Estimated Air Leakage (SCFM)
5	0.0005
20	0.007
35	0.012
45	0.023
60	0.034
75	0.046

Table 2. Air leakage at different air pressure.

Air Pressure (PSIG)	Estimated Air Leakage (SCFM)
50	0.001
200	0.005
350	0.009
450	0.014
560	0.023
700	0.035

The above results indicated that the preliminary results from prototype testing and computational simulation are very close, and both results verify the creditability and feasibility of this new magnetic energy-saving sealing system.

Conclusion

Currently the lubrication oil and gaseous media leakages are common and difficult engineering problems that affect the industrial machinery function and performance. This new magnetic energy-saving sealing system has been developed to reduce the oil and gaseous media leakages in industrial machinery. All the theoretical, mechanical and magnetic analysis, computational modeling and simulation, and prototype tests verify that this sealing system can significantly reduce the oil and gaseous media leakages in industrial machinery. Its sealing performance is reliable due to the firmly bonded and strong forces between the magnetic particles and reciprocating pistons/shafts. This seal mechanism is also durable if compared with regular seals including rubber seal, diaphragm seal, corrugated pipe seal. This new sealing technology brings energy-saving to the industrial machines because of higher machine efficiency with less frictional force between surfaces of seal and pistons/shafts and decreased internal gaseous media contamination. Furthermore this magnetic sealing product will contribute to the future exploitation, popularization, and application of the rich rare-earth elements/materials.

References

[1]. H. Quelle, G. González-Jorge, L. Domarco, Magnetic field cancellation on the surface of superconducting rings: Influence on the flux creep, *Cryogenics* 48 (2008), pp. 2-5.

[2]. H. González-Jorge, B. Liñares, I. Quelle, E. Carballo and G. Domarco, Measurement of critical current in superconducting rings, *Measurement* 40 (2007), pp. 378-380.

[3]. H. González-Jorge, J. Peleteiro, E. Carballo and G. Domarco, Inducting technique and trapped field in ring-shaped Superconductors, *Physica* 433 (2005), pp. 132-134.

[4]. U. Topal, L. Dorosiskii, H. Ozkan, and H. Yabuz, Effect of thermal neutron irradiation in boron-doped melt-textured YBCO superconductors, *Physica* 388 (2003), pp. 401-403.

[5]. L. Shlyk, G. Krabbes, G. Fuchs, K. Nenkov and B. Schupp, Flux pinning and magnetic relaxation in melt-processed $YBa_2Cu_3O_{7-d}$ dopped with Li, Journal of Applied Physica 96 (2004), pp. 371-374.

Design of An Intelligent Solar Tracking System Using PIC18F452 Micro Controller

Mostefa Ghassoul
mghassoul@eng.uob.bh
Department of Chemical Engineering
College of Engineering
University of Bahrain

Abstract:

This paper presents the design of a microchip PIC 18F452 micro controller based tracking solar energy system, where the extracted solar energy is optimized. The main advantage of the system is that it is programmed, so that it could rotate the solar panels at a pre-programmed angle and at a pre-determined periods by using short pulses. This is done by delivering programmed pulses which drive the rotating motors through the same programmed angle. The micro controller detects two light dependent sensors, one detecting the optimal inclination angle towards the sun, and the second guiding the solar panels to align themselves with it when required. This is done by measuring the voltage difference between the two sensors. If he difference is greater than the preset offset value, the panels are rotated so that they align themselves with the pilot. If the difference is less than the offset, then the panels maintain their current position while the pilot keeps tracking the sun. The choice of the flash micro controller is because it could be reprogrammed on-line using the technique known as In-Circuit Serial Programming as well as its very low price and size.

Keywords:

Panels, pilot, micro controller, LDR, sunrise, sunset.

Introduction:

Solar energy has been used for the last few decades and certainly, it will continue to be used extensively in the future, especially with the diminishing of alternative resources particularly carbon based energy resources such as oil, gas and coal. This has been encouraged by the improvement of the cells efficiency where it was merely 17% in the early 90's to pass the 30% mark lately. But, unfortunately, it is still suffering from few draw backs. The first one is the relatively high cost of the cells. The second one is that the efficiency is still low even at the current rate of 30%, and the third one is the positioning of the solar cells, especially during the long summer days where the sun shines for more than 16 hours a day and fixed cells do not extract maximum energy.

Currently, people using solar energy fix the panels mid way between the geographical east and west with approximately 30 degrees towards the south. Studies have shown that this is not ideal positioning in order to maximize energy extraction.

A better way is to continuously orient the panels towards the sun. This necessitates the continuous rotation of the later at a very low speed so that they lock towards the sun all the time. As it is known, at very low speed, the electric motors draw a very high current, so an important part of the extracted solar energy will be fed back to the motor and this not only results in the loss of most the energy but heats up of the motor as well. This renders the scheme inefficient.

An improved scheme(1) was implemented by continuously running a timer, delivering constant short bursts at predetermined periods distributed over the length of the day. These pulses drive the motor through pre determined angles, at constant intervals of time. Although this scheme has improved the efficiency considerably, it still suffers from a major draw back. It is not intelligent enough so that it only rotates the panels to the new position if they deliver higher energy. In fact, sometimes the driving motors consume more energy than the cells produce.

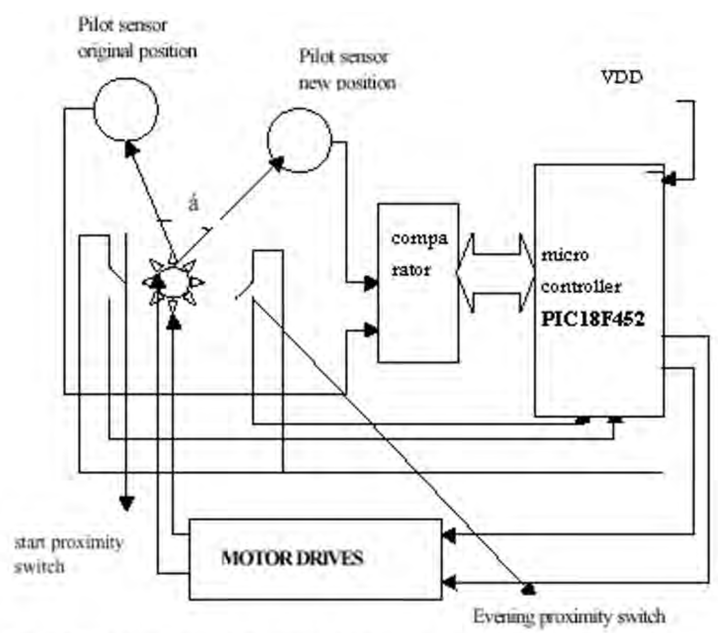
This paper discusses a better improvement to this scheme. It presents an optimal positioning mechanism of the panels so that maximum energy is extracted. The position is determined by using a search technique, using a pilot scheme. The pilot is built around a small sensor mounted on a miniature electric motor which only consumes a very tiny amount of energy. A comparison procedure compares between the energy of the panels and that of the pilot. If the latter is bigger, then the panels are aligned with it, otherwise they stay in their current position. The procedure repeats itself during the day. The tracking system is designed around a PIC 18F452 micro controller. The choice of this type of controller is due to the fact that it is cost effective, easily programmable, and has many features such as the analogue to digital conversion. It could also be erased and reprogrammed online using the In-Circuit Serial Programming technique (ICSP). It has also four timers ideal for this application.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_34, © Springer Science+Business Media B.V. 2010

System design:

The design depends on two light dependent sensors. One mounted on the solar panel and the second on the miniature motor (pilot). The miniature motor only draws a very small current, and is used to search for the position where maximum possible energy could be extracted. Before sunrise in the morning, both panels and pilot are oriented towards the geographical east (known as start run). For synchronization purpose, this position is detected by two sensors. A light dependent resistor (LDR) detects the sun rise and a proximity switch detects the position. On sun rise, the pilot starts rotating through a fixed angle determined by the micro controller internal timers. Two internal timers Timer0 and Timer1 in cascade are used. Timer0 is programmed as a pre- scalar dividing the system clock (500KHz obtained from the 4 MHz crystal) by 1:256, and Timer1 is used to count the clocks produced by the pre-scalar. When the time elapses, the timer delivers a programmed pulse proportional to the angle which the pilot is driven through to the new position, so that it keeps facing the sun. On the trailing edge of the pulse, the panels and pilot voltages are compared through a hardware comparator, and the output is used to trigger the controller accordingly. If the panels voltage is less than that of the pilot by the predetermined threshold, then the panels are moved to align themselves with the pilot. If this condition is not met, then the processor records the total number of pulses by which the panels lag the pilot so that when the condition is met, the panels are rotated through an angle equivalent to the total number of missed pulses in order to align themselves with the pilot (the duration is determined using the cascaded timers mentioned earlier). This procedure repeats itself with each driving pulse until one of the following conditions is met. Either the end proximity switch is activated or the sun sets activating the pilot LDR. In the first case, the proximity switch is oriented towards the west. During longer days of the year, it makes sure that the panels do not over run their trajectory. If this is the case both panels and pilot stop and wait for the sun to set which is detected by the LDR. On sun set, the system rotates back all the way to the starting position. When the start proximity switch is activated, the system comes to a halt (facing geographical east) and waits for the sun rise of the next day to repeat the procedure.

A block diagram of the tracking system is shown in figure (1).

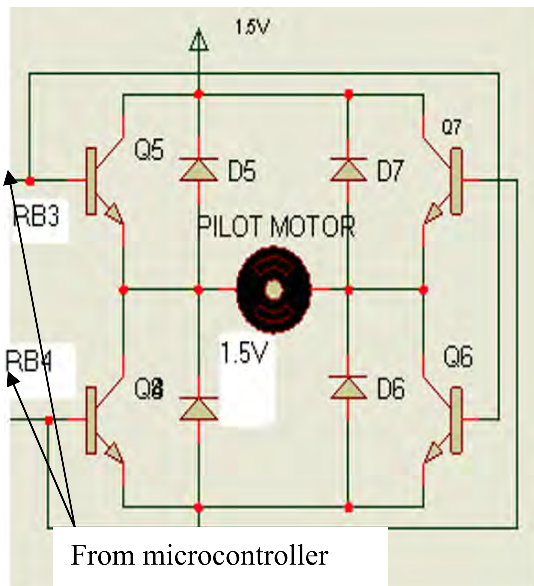


Figure(1): tracking system to determine the maximum sun inclining angle

Motor drive:

The drives for pilot and panels motors are identical except the amount of power delivered to each of them. The only difference is that the pilot is driven by the micro controller through its output port B pin RB3 for the forward rotation and RB4 for return rotation whereas the panels are driven through pin RB5 for the forward rotation and RB6 for the return rotation. When RB3 is activated, Q5 and Q6 are both closed, and the pilot motor is energized, and rotates in clockwise direction. When RB4 is high, Q7 and Q8 are energized and the pilot motor rotates in counter clockwise direction. Unlike RB3, RB4 is connected to a set software flip/flop (refer to figure2) which drives the system back to the origin by hitting a proximity switch which resets the flip flop and the system comes to a stand still.

The same procedure takes place with the panels. When RB5 is activated, they rotate forward and when RB6 is activated, they rotate backwards. As it is well known, the flywheel diodes are used to protect the micro controller from any current surge due the coil effects (refer to figure (3).



Fig(2) Pilot motor control

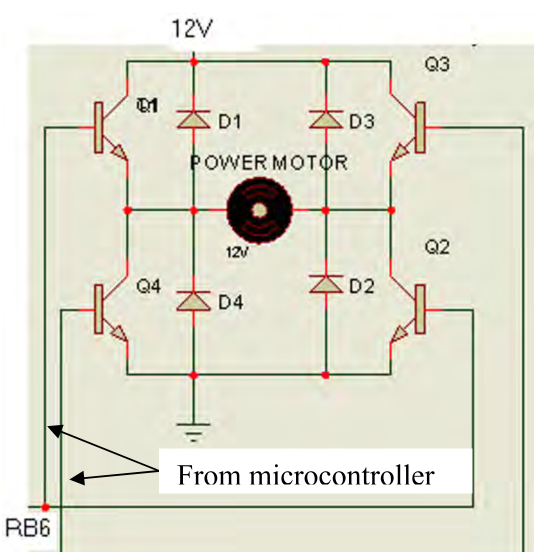


Figure (3): power drive of panel motor

Sun rise and sun set capture:

The sun rise and sun set detection is straightforward. It is done through a light dependent resistor as shown in figure (4). When sun rises, the LDR resistance decreases and the current through it increases. The voltage at the transistor base builds up until the transistor is forced into saturation. The collector current in turn increases and the relay closes connecting the supply to the micro controller input PORTA pin RA0. When darkness falls, the LDR resistor increases and the transistor base current decreases cutting the collector current, and in turn disconnecting the relay. The flywheel diode is connected to protect the transistor against the di/dt effect. An identical circuit is used for the panels but the input is through controller pin RA3

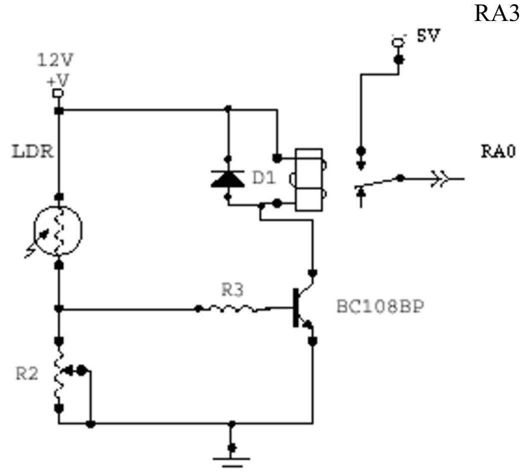


figure 4 sun rise and fall capture

Voltage level detection:

The voltage difference between that of the panels and that of the pilot is detected using the circuit shown in figure (5). The MAIN CELL LDR is connected to the panel and the PILOT CELL LDR is connected to the pilot. The two variable resistors, at the bottom of the figure, are used so that the voltage difference could be controlled. If the panels voltage is higher, the output of the comparator is high. This in turn activates the micro controller to rotate the panels so that they will be aligned with the pilot. If the pilot voltage is higher, then the panels are left standstill as explained earlier.

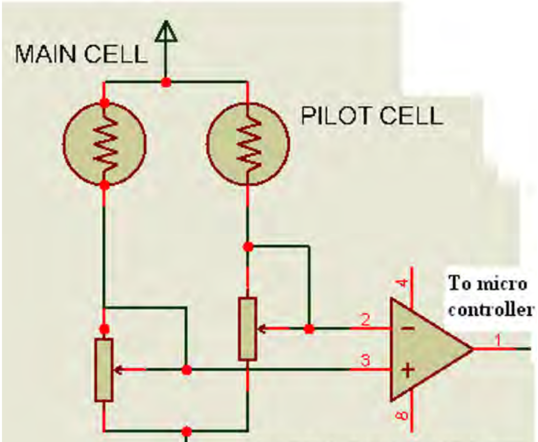


Figure 5: Detection of voltage difference between pilot LDR and cells LDR

Chronology of the system:

Figure (6) shows the itinerary of the system trajectory during the course of the day. The first pulse shows the correct orientation of the pilot and panels towards the east waiting for sun rise. When the sun rises (sun rise), the pilot sensor is activated and it starts rotating at a pulsating rate(say 2seconds on and half and an hour off. This is software programmable) as explained earlier. On each falling edge, the comparison process takes place. When the pilot voltage is smaller than or equal to that of the panels, the required pulses to drive the panels motor are not generated (panels do not follow the pilot). And when the pilot voltage is bigger than that of the panels by the programmed offset, the solar panels follow the pilot by the same number of missed pulses (solar panel following the pilot). When the pilot hits the end proximity switch, the system stop and waits for sun set. When this happens, the system rotates back to the start and waits for the following morning.

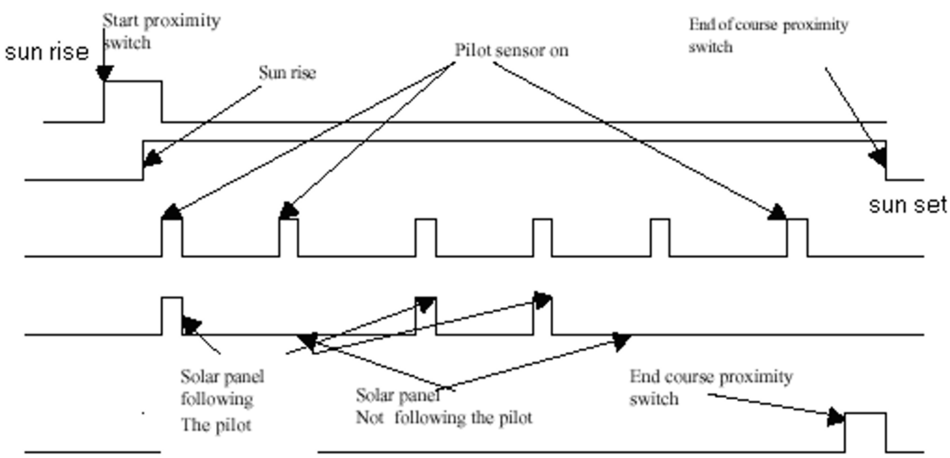


figure6: chronology of the pilot and panels movement

Software

The program was written using the microchip assembly language. It was tested using the MPLAB utility and using the ICD2 emulator.

Test and conclusions:

The prototype is tested in the laboratory using a mobile lamp as a light source, a pilot and a single panel delivering 24 volts. At the beginning, the prototype is oriented towards the east where the proximity switch (start proximity switch of Omron M8 eight millimeter spacing type mounted on the panel pivot refer to figure (1)) is closed forcing the pin RA3 of micro controller input PORTA to the high state, forcing in its turn the motors (panel and pilot) to the stand still state.

When the lamp is turned on, the timer starts and delivers pulses of one millisecond every five minutes interval. The pilot sensor rotates through an angle of approximately twenty degrees for each pulse.

On the pulse trailing edge, the pilot and the panels voltages are compared. Because the two are still close to each other, their voltage difference is still less than the offset.(the offset is set using the two potentiometers as to 2V shown in figure(5)). As a result, the panel maintains its current position. This procedure repeats itself for the next three pulses. On the fourth one, the voltage difference is more than the offset and the panel rotates for approximately eighty degrees to align itself with pilot. This is equivalent to the four missed pulses.

After 180 degrees, the pilot activates the end proximity switch, which in turn interrupts the controller to drive the system to standstill. The lamp is moved manually When the lamp is turned off, the system rotates back to its initial state.

As a conclusion, a cost effective intelligent sun tracking system to extract maximum solar energy possible was designed. Unlike what was reported in literature, the main advantage of the system is that it is intelligent enough so that the panels track the sun only if that contributes to extra energy extraction and at the same time, the energy consumed by panels driving motor is less than that extracted. The system can also align itself to perfection either on sunrise or sunset so no drift is allowed. Another main advantage is that it takes the days variation during the year into consideration so to make sure that the panels do not lead the sun during longer days nor lag it during shorter ones otherwise the system will completely go out of synchronization.

Finally It is hoped that the system could be further looked at for commercial exploitation.

References:

[1] "A programmable logic system to control a solar panel to pump water from a well" Ghassou M. and Radwan F., 5th IEEE TEM, KFUPM Dhahran, pp 67-70

[2] "Design of PLC based tracking system to maximize solar energy conversion" Ghassoul, M., proceedings of the 8th IEEE TEM, KFUPM Dhahran,17-18 april 2001

[3] "28/40 pin 8 bit CMOS flash micro controllers", microchip datasheet 30292, 1998

[4] "RC model aircraft motor control", publisher J.,microchip technology INC, 2002

[5] "Maximum power solar converter" Application AN1211, Charais,J. Microchip Technology, 2008

[6] "Low power capacitive sensing with the capacitive sensing module" Application AN1103, Floricica,A. Microchip Technology, 2008

[7] "Software handling for capacitive sensing" Ferme,T. Microchip Technology, 2007

[8] "Speed control of three phase induction motor using PIC 18 micro controller" Yedamale P., Microchip Technology INC, 2002

[9] "Brush DC servomotor implementation using PIC 17C756A", Benling S., Microchip Technology INC,1999

[10] "Servo control of a DC brush motor", Bucella T., Teknic INC,1997

[11] "Programmable logic controllers", Par E.A., 1996, published by Newnes

[12] "Programmable logic devices and logic controllers", Manlado E. and Co, published by Prentice Hall, 1997

Adaptive Two Layers Neural Network Frequency Controller for Isolated Thermal Power System

Ognjen Kuljaca, Krunoslav Horvat*, Jyotirmay Gadewadikar
Systems Research Institute, Department of Advanced Technologies, Alcorn State University
1000 ASU Drive #360, Alcorn State, MS 39096, USA
okuljaca@alcorn.edu, jyo@alcorn.edu
*Brodarski Institute
Av. V.Holjevca 20, Zagreb, HR 10020, Croatia
kruno@hrbi.hr

Abstract: **- An adaptive neural network control scheme for thermal power system is described. Neural network control scheme presented in this paper does not require off-line training. The online tuning algorithm and neural network architecture are described. The performance of the controller is illustrated via simulation for different changes in process parameters and for different disturbances. Performance of neural network controller is compared with conventional proportional-integral control scheme for frequency control in thermal power systems. Neural network control scheme described here is not linear-in-parameter. Neural network has two layers and nodes weights in both layers are tuned.**

I. Introduction

The paper deals with the neural network (NN) frequency controller for isolated thermo power system. In modern power systems stable frequency is one of the main measures of power quality. Frequency control becomes more and more important as power systems enter the era of deregulation. Frequency of the system changes if change in load is not met with the adequate change in generation. It is becoming very hard, if not impossible to schedule loads precisely, thus the load fluctuations in the power system are becoming more explicit. Also, it is becoming very hard to have determined and constant primary controllers and turbines. Emerging markets of ancillary services mean that primary controllers and turbines that are used in secondary control change constantly. These changes can cause serious problems when conventional control schemes are used, including problems with stability, unless the primary turbines and controllers are carefully selected. To avoid possible instability with changes of parameters in the system, conventional secondary controllers are implemented using smaller integral gains than optimal performance would require.

The literature about frequency and load – frequency control is numerous ([1], [2], [3], [4], [5], [6], [7], and many others). Many non-adaptive schemes are given in ([1], [2], [3], [4], [5] and [6]. However, as described before, the modern power systems in deregulated environment are subject to frequent parameters changes which may diminish the quality of control when nonadaptive controllers are used.

It is thus natural to examine adaptive control schemes. Conventional adaptive control schemes, like one described in [7] can perform well. However, they require parameter identification and with fast and constant changes this could pose problems in controller realization. Thus neural network control may be employed with no need of constant parameters identification. NN load-frequency control is described in [8], [9] and [10]. The results obtained by using NN controllers are good. However, the described controllers require training. We provide here a performance analysis of adaptive NN controller that does not require training. The neural network is capable of on-line learning. The controller described here is an improved version NN control scheme given for the first time in [30]. Unlike neural network controller given, our scheme here is augmented in such way that both layers of neural network are tunable, not only output layer like in [30].

The paper is organized as follows. In Section II. are given some mathematical preliminaries. Model of isolated thermo power system is given in Section III. Neural network control scheme is described in Section IV. In Section V. the simulation results are given and in Section VI. is given the conclusion.

II. Mathematical Preliminaries

Let R denote the real numbers, R^n the real n-vectors, R^{mxn} the real mxn matrices. Let S be a compact simply connected set of R^n. With map $f: S \to R^m$, let us define $C^m(S)$ the space such that f is continuous. Let $\|\bullet\|$ be any suitable vector norm. The supremum norm of $f(x)$ over S is defined as:

$$\sup\|f(x)\|,\ f: S \to R^m,\ \forall x \in S \tag{1}$$

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_35,

Given $x \in R^{N_1}$, a two-layer NN (Fig. 1) has a net output given by

$$y = W^T \sigma(V^T x) \qquad (2)$$

where $x = [1 \quad x_1 \quad \cdots \quad x_{N_1}]^T$, $y = [y_1 \quad y_2 \quad \cdots \quad y_{N_3}]^T$ and $\sigma(\bullet)$ the activation function. If $z = [z_1 \quad z_2 \quad \cdots]^T$, we define $\sigma(z) = [\sigma(z_1) \quad \sigma(z_2) \quad \cdots]^T$. Including "1" as a first term in $\sigma(V^T x)$ allows one to incorporate the thresholds as the first column of W^T. Then any tuning of NN weights includes tuning of thresholds as well.

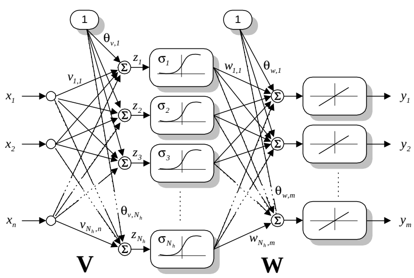


Fig. 1: Two layer neural network

The main property of NN we are concerned with for control and estimation purposes is the function approximation property ([16], [20]). Let $f(x)$ be a smooth function from $R^n \to R^m$. Then it can be shown that if the activation functions are suitably selected, as long as x is restricted to a compact set $S \in R^n$, then for some sufficiently large number of hidden-layer neurons L, there exist weights and thresholds such that

$$f(x) = W^T \sigma(V^T x) + \varepsilon(x) \qquad (3)$$

The value of $\varepsilon(x)$ is called the neural network functional approximation error. In fact, for any choice of a positive number ε_N, one can find a neural network such that $\varepsilon(x) \le \varepsilon_N$ for all $x \in S$.

III. Isolated Thermopower System

The model of an isolated thermo power system is shown in Fig. 2.
The transfer functions in the model are:

$$G_g = \frac{1}{1 + s \cdot T_g}, \qquad (4)$$

$$G_t = \frac{1}{1 + s \cdot T_t}, \qquad (5)$$

$$G_s = \frac{K_s}{1 + s \cdot T_s}, \qquad (6)$$

where G_g, G_t and G_s are representing turbine governors, control turbines and the power system respectively. Parameters are: T_g – turbine governor time constant, T_t – turbine time constant, K_s – power system gain, T_s –power system time constant. Such models are described in more details in [1], [2], [3], [4], [5], [6], [7], and many others.

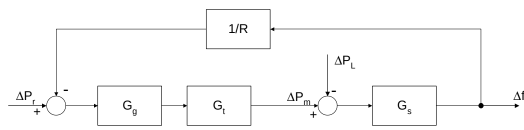


Fig. 2: The model of isolated thermo power system

It is also shown that the system in Fig. 2 is always asymptotically stable if R is positive number. In real systems that is always the case. The system is linear and the need for adaptive control or use of the function approximation property of the neural network is not obvious since there are no nonlinearities in the controlled plant.

However, all the parameters can and do change during the operation. Thus, it is conceivable that adaptive control scheme would perform better than nonadaptive control.

The most usual way of control is to use linear PI controllers. The controller in that case takes the change of power system frequency Δf as the input and produces the control output ΔP_r as output. That signal is fed to turbine governor in order to counter the changes caused by the change in the load ΔP_L. The turbine output is the mechanical power ΔP_m. However, introduction of integral action also means that the system can become unstable.

The system shown in Fig. 2 can be represented in state space as

$$\dot{x} = \begin{bmatrix} -\frac{1}{T_g} & 0 & -\frac{1}{RT_g} \\ \frac{1}{T_t} & -\frac{1}{T_t} & 0 \\ 0 & \frac{K_s}{T_s} & -\frac{1}{T_s} \end{bmatrix} x + \begin{bmatrix} \frac{1}{T_g} & 0 \\ 0 & 0 \\ 0 & -\frac{K_s}{T_s} \end{bmatrix} \begin{bmatrix} \Delta P_r \\ \Delta P_L \end{bmatrix} \qquad (7)$$
$$= Ax + Bu$$

The state vector x is

$$x = \begin{bmatrix} y_g \\ \Delta P_m \\ \Delta f \end{bmatrix} \qquad (8)$$

where y_g is output of the turbine controllers. These states are physically available, so this representation will allow for the NN control scheme design.

IV. Adaptive Neural Network Control

The neural network is build as shown in Section 2. The first layer weights, sigmoid functions, are initialized randomly and then fixed to form basis φ(x). The NN output is

$$y = W^T\sigma(V^T x) \tag{9}$$

This architecture is an adapted form of the tracking NN controller described in [23], [24], [25], [26], [27] and numerous other papers. However, there are some differences. Namely, here the problem is control, not tracking. Also, there is no special robustifyng term or PD controller parallel to neural network. Since there are significant time constants present in the controlled plant, derivative part will not have an effect. Proportional part is not needed to initially stabilize the system since uncontrolled system is always stable. Proportional gain K was still used in scheme given and analyzed in [30], but simulation analysis had shown that proportional gain does not improve the performance of the control scheme. At last, unlike in papers mentioned above, we don't use filtered error approach.

It is assumed that the load disturbance ΔP_L is bounded so that

$$\Delta P_L \le \Delta P_M \tag{10}$$

Since our neural network acts like tunable gain (no dynamics), (10) means that state vector x is restricted to a compact set.

This assumption is always true as long as we deal with the power system in the normal mode of operation. If the load disturbance is too big there cannot be any control action since in that case the system just doesn't have enough power generation capability available. The protection functions take over in that case and some loads have to be disconnected.

Let us define vector z as:

$$z = \begin{bmatrix} 1 \\ x \end{bmatrix} \tag{11}$$

Let the control signal be given by

$$\Delta P_r = W^T\sigma(V^T z) \tag{12}$$

and the weight updates are provided by

$$\dot{W} = F\sigma(V^T z)\Delta f - k_w \|z\| F W. \tag{13}$$

$$\dot{V} = Gz(\sigma(V'^T z)^T W \Delta f)^T - k_v \|z\| G V. \tag{14}$$

These update laws where obtained from Lyapunov like stability analysis of the system as described in [25] and [28].

V. Simulation Results

The simulations were performed for the following set of parameters: T_q = 0.08 s, T_t = 0.3 s, K_s = 120 pu/s, R = 2.4 Hz/pu. The neural network had 20 nodes in the hidden layer; initial network values were initialized to small random numbers. The neural network parameters were $k_v=10^{-2}$, $k_w=10^{-3}$, G=diag(7) and F=diag(0.05). The network used sigmoid activation function. The responses of the system with NN control is compared with the usual PI controlled system with proportional gain of the controller k_p = 0.08 and integral gain k_i = 0.6.

The load disturbance was simulated as $\Delta P_L = 0.1$ p.u. and $\Delta P_L = 0.1\sin(2\Pi f_L t)$ p.u. Amplitude of 0.1 p.u for disturbance presents the disturbance of 10% of available generation and it is quite big in practical situations.

The simulation results for different frequencies of disturbance are shown in Fig. 3 – Fig. 10.

Fig. 3 depicts the responses of the system controlled with PI controller and system controlled with neural network controller for step disturbance at 0 s and 500s. The details for this situation are shown in Fig. 4 and Fig. 5. It can be seen that in case of step disturbance NN controller settles faster and has smaller steady state error.

Fig. 6 – Fig. 9 give the response for different disturbance frequencies. It can be seen that NN control scheme outperforms conventional PI controller in all cases. Frequency deviation is two or three times smaller when NN controller is used.

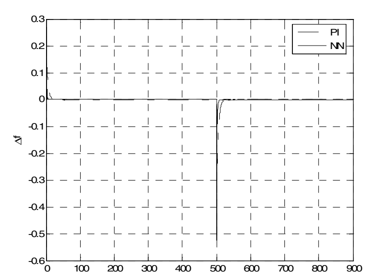


Fig. 3: The step input P_L=0.1 p.u.

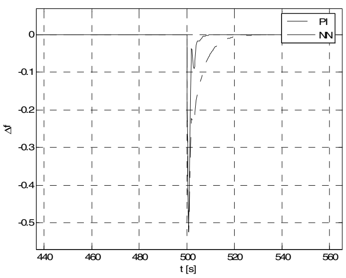


Fig. 4: The step input P_L=0.1 p.u. - detail

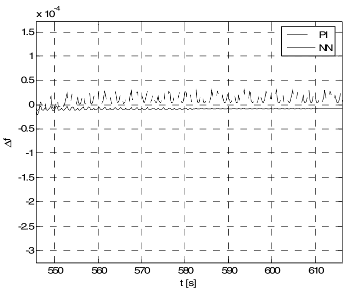


Fig. 5: The step input P_L=0.1 p.u. - detail

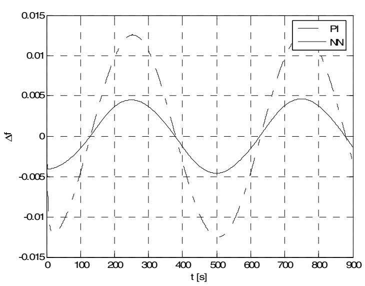


Fig. 6: The responses for f_L=0.002

The response of the system for changed turbine governor and turbine parameters (T_g=0.09s and T_t=0.8s) is shown in Fig. 6 and Fig. 7. Again, NN control outperforms PI controller.

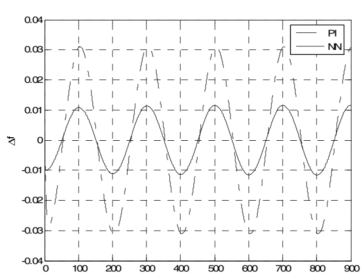


Fig. 7: The response for f_L=0.005

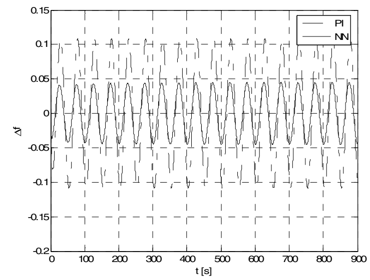


Fig. 8: The response for f_L=0.02

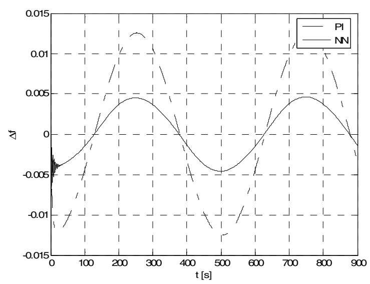


Fig. 9: The response for f_L=0.002 and changed governor and turbine parameters

VI. CONCLUSION

The results of an initial research in neural network control of power systems with two layers nonlinear in parameter neural network controller are shown. The simulation results show that controller performs well and adapts to changing parameters. The controller does not require off-line training phase. Analysis of responses shows that neural network control scheme outperforms conventional control scheme. By defining the neural network differently, having output layer weights W defined as a matrix this scheme can be adjusted to deal with the multivariable systems. The performance analysis here shows that is would be worth to continue with the research effort toward two tunable layer neural network controllers for interconnected systems as well as for the systems with generation rate constraint.

REFERENCES

[1] Elgerd, O.I. "Control of Electric Power Systems", IEEE Proceedings, pp. 4-15, 1981

[2] Elgerd, O.I., Fosha, C.E., "Optimum Megawatt-Frequency Control of Multiarea electric Energy Systems" IEEE Transactions Power Apparatus and Systems, PAS-80, No. 4, pp. 556-562, 1970

[3] Fellachi, A., "Optimum Decentralized Load Frequency Control" IEEE Transactions on Power Systems, Vol. PWRS-2, No. 2, pp. 379-384, 1987

[4] Fosha, C.E., Elgerd, O.I., "The Megawatt-Frequency Control Problem: A New Approach Via Optimal control Theory" IEEE Transactions Power Apparatus and Systems, PAS-80, No. 4, pp. 563-577, 1970

[5] Hamza, M.H., Agathoklis, P., Chan, W.C., "A Combined Selftuning and Integral controller for Load Frequency Control of Interconnected Power systems "Regelungstechnik 30. Jahrgang Heft 7 pp. 226-232, 1982

[6] Kundur, P., Power System Stability and Control. McGraw-Hill, Inc.,1993

[7] Liaw, C.M., "Design of a Reduced-Order Adaptive Load-Frequency Controller for an Interconnected Hydrothermal Power System "International Journal of Control, Vol. 60, No. 6, pp. 1051-1063, 1994

[8] Demiroren, A., Zeyneigly, H.I., Sengor, N.S., "The application of ANN technique to load-frequency control for three-area power system", 2001 IEEE Porto Power Tech Conference, Porto, Portugal

[9] Bevrani, H. "A novel approach for power system load frequency controller design", Transmission and Distribution Conference and Exhibition 2002: Asia Pacific. IEEE/PES , Volume: 1 , 6-10 Oct. 2002

[10] Birch, A.P., Sapeluk, A.T., Ozveren, C.T., "An enhanced neural network load frequency control technique", International Conference on Control, 1994. Control '94. Volume 1, 21-24 Mar 1994

[11] R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," IEEE Trans. Info. Theory, vol. 39, no. 3, pp. 930-945, May 1993.

[12] A.G. Barto, "Reinforcement learning and adaptive critic methods," Handbook of Intelligent Control, pp. 469-492, Van Nostrand Reinhold, New York, 1992.

[13] J. Campos and F.L. Lewis, "Adaptive critic neural network for feedforward compensation," Proc. American Control Conf., vol. 4, pp. 2813 –2818, San Diego, 1999

[14] F.-C. Chen, and H. K. Khalil, "Adaptive control of nonlinear systems using neural networks," Int. J. Contr., vol. 55, no. 6, pp. 1299-1317, 1992.

[15] M. J. Corless and G. Leitmann, "Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems," IEEE Trans. Automat. Contr., vol. 26, no. 5, pp. 850-861, 1982.

[16] G. Cybenko, "Approximation by superpositions of a sigmoidal function," Math. Contr. Signals, Syst., vol. 2, no. 4, pp. 303-314, 1989.

[17] C. A. Desoer and S. M. Shahruz, "Stability of dithered nonlinear systems with backlash or hysteresis," Int. J. Contr., vol. 43, no. 4, pp. 1045-1060, 1986.

[18] B. Friedland, Advanced Control System Design, Prentice-Hall, New Jersey, 1996.

[19] B. Igelnik and Y. H. Pao, "Stochastic Choice of Basis Functions in Adaptive Function Approximation and the Functional-Link Net," IEEE Trans. Neural Networks, vol. 6, no. 6, pp. 1320-1329, Nov. 1995.

[20] K. Hornik, M. Stinchombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, pp. 359-366, 1989.

[21] B.S. Kim and A.J. Calise, "Nonlinear flight control using neural networks," AIAA J. Guidance, Control, Dynamics, vol. 20, no. 1, pp 26-33, 1997.

[22] J.-H. Kim, J.-H. Park, S.-W. Lee, and E. K. P. Chong, "Fuzzy precompensation of PD controllers for systems with deadzones," J. Int. Fuzzy Syst., vol. 1, pp. 125-133, 1993.

[23] S.-W. Lee and J.-H. Kim, "Control of systems with deadzones using neural-network based learning control," Proc. IEEE Int. Conf. Neural Networks, June 1994, pp. 2535-2538.

[24] F. L. Lewis, C. T. Abdallah, and D. M. Dawson, Control of Robot Manipulators, Macmillan, New York, 1993.

[25] F. L. Lewis, A. Yesildirek, and K. Liu, "Multilayer neural-net robot controller with guaranteed tracking performance," IEEE Trans. Neural Networks, vol. 7, no. 2, pp. 1-11, Mar. 1996.

[26] F. L. Lewis, K. Liu, and A. Yesilidrek, "Neural net robot controller with guaranteed tracking performance," IEEE Trans. Neural Networks, vol. 6, no. 3, pp. 703-715, 1995.

[27] F. L. Lewis, K. Liu, R. R. Selmic, and Li-Xin Wang, "Adaptive fuzzy logic compensation of actuator deadzones," J. Robot. Sys., vol. 14, no. 6, pp. 501-511, 1997.

[28] F.L. Lewis, S. Jagannathan, and A. Yesildirek, Neural Network Control of Robot Manipulators and Nonlinear Systems, Taylor and Francis, Lindon, 1998.

[29] N. Sadegh, "A perceptron network for functional identification and control of nonlinear systems," IEEE Trans. Neural Networks, vol. 4, no. 6, pp. 982-988, Nov. 1993.

[30] Kuljaca, O., Lewis, F. Tesnjak, S., "Neural Network Frequency Control for Thermal Power Systems", IEEE Control and Decision Conference, 2004

Optimal Controller Comparison Using Pareto Fronts

David Moore
Department of Electrical Engineering
University of Cape Town
Rondebosch
South Africa

***Abstract* – New design methods in Control System are regularly proposed. These new methods are typically compared to existing methods in a focussed manner than highlights certain criteria, but may neglect other criteria. This paper deals with the use of Level Diagram visualisation for Pareto Fronts as a device for comparing modern control schemes in a robust and objective manner. Specifically it uses a comparison of two methods: PI and Sliding Mode Control as a test case.**

***Keyword* – Multiobjective Optimization, Differential Evolution, Pareto Front, Level Diagrams, Sliding Mode Control.**

I. Introduction

New methods are constantly being proposed for the design of Control Systems (e.g. VSC in 1970s, H-infinity in the 1980s, etc). More often than not, the new methods are compared to previous ones in a focused manner that highlights one or two criteria. This can be misleading in some cases [l] and a more comprehensive comparison is required.

This paper presents a framework for comparing Control design methods using high dimensional Pareto fronts, that are computed using Multi-Objective Optimization (MOO) 2] and investigated using the novel Level Diagram [3] visualisation. Specifically it compares the Sliding Mode Control 4] and PI Control methods in terms of many criteria.

To ensure that the comparisons are unbiased, the framework uses the best controllers that each method can produce. The performance of the controllers is quantified by cost functions (or costs) that are assumed to identify the best controllers. These costs are frequently and typically mutually exclusive.

Multi-Objective Optimization methods are utilized to find optimal (best) solutions to a given problem [2] where optimal is defined in terms of the trade-offs between computed cost functions and the concept of Pareto Efficiency 2].

A Pareto Efficient or Optimal situation is one where any change to make a member better off is impossible without making some other member worse off [2]. Solutions that are Pareto Efficient are described as non-dominant 5][6].

Many different types of MOO methods exist; they fall into four groups depending on when the decision maker adds preference information. These four groups are priori (before), progressive (during), posteriori (after), and no preference 7].

The framework proposed in this paper aims to make as broad an analysis of controller applicability to a specific problem as possible. Hence it uses a posteriori method, Differential Evolution (DE) [8], where as little subjective inference as possible is added before the decision maker is presented with information. DE is a stochastic population-based multi-objective optimizer. It uses real numbers and the evolution based ideas of crossbreeding and mutation to find solutions in the form of sets that are mutually optimal. Such solutions are typically represented as Pareto Fronts.

Pareto Fronts are traditionally visualised on a 2-D plot of cost vs. cost showing the trade-offs that can be made between various costs 2]. When the front exists in more than 2 dimensions it can become difficult to analyse [1] and even more difficult to visualize.

Level Diagrams 3] are a novel method for visualization of high dimensional Pareto Fronts that overcomes this limitation.

Most MOO are coupled to a decision maker and attempt to find a single or directed set of solutions to a given problem [2] rather than presenting a visualization of the trade-offs.

II. Method Overview

The success of a control method is typically dependent on the specifics of its design and implementation. It makes sense to compare two methods only when both are very well designed.

In order to ensure this, optimization is used to find the best controller that a method can produce, based on the costs that have been defined for the problem.

A. Multiobjective Optimization

Multiobjective Optimization finds the best costs by simultaneously optimizing two or more conflicting objectives subject to certain constraints 2]. It can be formalized mathematically as follows

1

$$J(\theta) = [J_1(\theta), \ldots, J_k(\theta)].$$

$$min_{\theta \in \Omega} J(\theta)$$

where θ is the input or decision vector, **Ω** is the decision space and $J(\theta)$ is the cost or objective vector. This

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_36,

optimization can then be explained as simultaneously minimizing all objectives of $J_i(\theta)$ for every J_i.

Multi-Objective Optimization yields a set of mutually optimal solutions Ω_P. It is unique for a given set of costs. This optimal set is commonly referred to as the Pareto Front [5].

i. Differential Evolution

Evolutionary Algorithms (EA) are a type of optimization method increasingly being used in engineering design and practice 5]. There are many different EA methods available, each with different characteristics. One such method is Differential Evolution, which is a popular method for multi-objective optimization problems [5].

The Pareto Differential Evolution Algorithm 6] is a population-based optimizer, that finds a population of *n* members making up a discrete approximation Ω_P^* of the Pareto front Ω_P.

ii. 2-D Pareto Front

2-D Pareto Fronts are a representation method used to compare trade-offs between optimal solutions to multi-objective optimization problems, based on their costs.

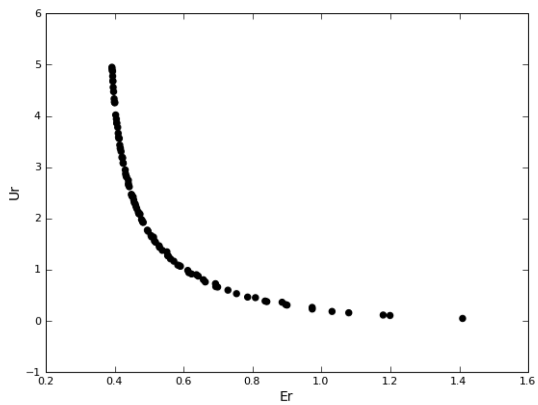


Figure 1: 2-D Pareto Front

An example of a 2-D Pareto Front is given in Figure 1, in this case showing a discrete approximation of the Pareto Front Ω_P^* for two costs.

B. Level Diagrams

These visualisations [3] are used for demonstrating higher dimensional Pareto fronts (fronts of dimension greater than 3) such that useful information can still be visually interpreted from them.

In Level Diagrams each cost is classified based on it's proximity to an ideal point. Each cost is first normalised with respect to its minimum and maximum values, such that each has equal importance.

$$J_i^{max} = \max_{\theta\in\Omega_P^*} J_i(\theta)\,,\; J_i^{min} = \min_{\theta\in\Omega_P^*} J_i(\theta)\,,\; i = 1,\dots,k \quad (1)$$

$$\hat{J}_i = \frac{J_i(\theta)-J_i^{min}}{J_i^{max}-J_i^{min}} \rightarrow 0 \le \hat{J}_i(\theta) \le 1 \quad (2)$$

A norm is then applied to the normalised costs to evaluate the distance to the ideal point. The Euclidean norm (3) gives the best conventional geometric representation of the shape of the Pareto Front [HYPERLINK \l "XBl08" 3] and is used in this paper.

$$\|\hat{J}(\theta)\|_2 = \sqrt{\sum_{i=1}^{k} \hat{J}_i(\theta)^2} \quad (3)$$

Each cost (J_i) and input (θ_i) is plotted on a separate set of axes. They are plotted on the X axis against their corresponding norm $\|\hat{J}(\theta)\|_2$ on the Y axis.

This effectively places all corresponding information for any J_i or θ_i at the same position on the Y axis.

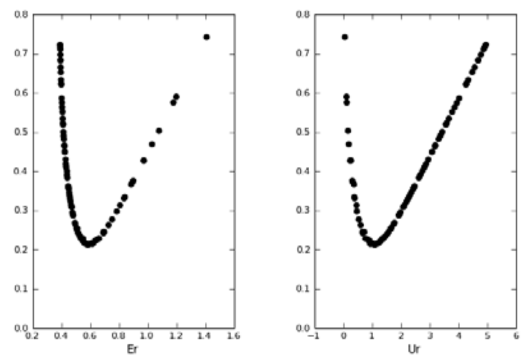


Figure 2: Level Diagrams

Figure 2 gives the Level Diagrams for the same set given in Figure 1.

An example of Level Diagrams with higher dimension, (in this case 7 costs) is shown in Figure 7 and Figure 8.

C. Proportional Integral Control

The two-term PI controller is an industry standard control method developed in the 1940s. A complete analysis and development of design techniques is given in [9].

D. Sliding Mode Control

Sliding Mode Control (SMC) is a robust non-linear control technique [4][10]. It is based on the concept that it is easier to control 1st-order systems than nth-order systems. The technique works by defining a sliding surface (4), (sliding mode, manifold or hyper-surface) in the state-space:

$$s = \left(\frac{d}{dt} + \lambda\right)^{n-1} \tilde{x} \quad (4)$$

This surface is defined such that the system trajectory has desirable dynamics when on the surface.

And the scheme is designed such that the feedback forces the system trajectory onto the sliding surface [4] in a reaching motion. It then remains on the surface through input switching in a sliding motion.

A reasonable modification [4] includes a $-\Phi s$ term that allows direct specification of both reaching and sliding rate dynamics.

$$u = u_{eq} - K sgn(s) - \Phi s \quad (5)$$

$$sgn(s) = \begin{cases} -1, & s < 0 \\ 1, & s > 0 \end{cases} \quad (6)$$

SMC is meant to be insensitive to parameter changes and be able to completely reject matched disturbances [11] due to the non-linear term $-K sgn(s)$.

Complete SMC design methods are presented in both [10][4].

One of the inherent difficulties of SMC is chatter that is induced by non-ideal switching. Minimization of this chatter is one of the primary focuses of research on SMC. A review of methods for reducing chatter is presented in [11].

For this paper a pragmatic engineering approach, called Integrated SMC, was used to deal with chatter. The method works by augmenting the SMC controller with an integrator on its output. The output of this integrator is the manipulated variable for the plant. There are three main reasons for the integrator's inclusion: Firstly, SMC design techniques call for systems with greater than first order dynamics, so first order systems need augmenting. Secondly, the integrator acts as a filter, removing the need for any other chatter correction. And finally it eliminates the finite offset error that would occur with first order plant models due to input disturbances.

One expected result from this modification is that integrating the SMC output reduces the controller's effectiveness at rejecting disturbances due to the delayed response that the integrator introduces.

III. Comparison Definitions

As stated above, PI and SMC control are compared to one another using Level Diagrams that are computed by the Pareto Differential Evolution Algorithm. Furthermore, a set of PI controllers that cancel the plant poles and lead to poor internal performance [12] was included for extra comparative purposes. Figure 3 shows the general plant setup with setpoint *r*, input *v* and output *d* disturbances.

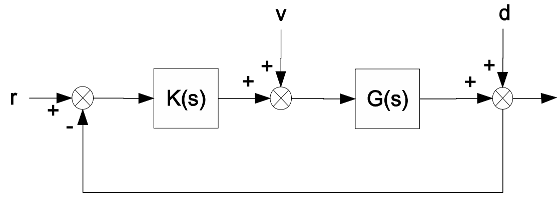


Figure 3: General Control Setup

Some definitions of the simulation are given below.

A. Plant Definition

The test plant is a small sized DC motor that represents a typical first order dynamic system with transfer function:

$$G_{motor}(s) = \frac{2}{1+8s} \quad (7)$$

The motor speed is regulated by the PI and (Integrated) SMC controllers.

B. Cost Definitions

The Pareto Differential Evolution algorithm works by optimizing the outputs (i.e. costs) for a given set of inputs (i.e. decision variables). The choice of these costs affects the outcome of the optimization.

In order to keep the comparison as general as possible, the costs were chosen to be as application agnostic as possible.

The first three costs are based on the Integral Square Error (9) of the plant error *e* (8), resulting from changes in the setpoint *r*, input *v* and output *d* disturbances introduced to the system.

$$e = r - y \quad (8)$$

$$ISE_{r,v,d} = \int_{t_{perturb}}^{t_{final}} e^2 dt \quad (9)$$

Similarly, the next three costs are based on the Integral Square Error for the plant control *u* (ISU). It is modified such that the resulting cost is a finite integral. This is done by subtracting *u*'s final value from *u* at each time step.

$$ISU_{r,v,d} = \int_{t_{perturb}}^{t_{final}} (u - u_{\infty})^2 dt \quad (10)$$

Both the $ISE_{r,v,d}$ and $ISU_{r,v,d}$ costs are abbreviated to $E_{r,v,d}$ and $U_{r,v,d}$ in the diagrams that follow.

The last cost included is a rough measure of the robustness of the controlled system to model changes. It finds the largest positive percentage change for each of the 6 other costs when the plant parameters A and T are varied by 50%. In order for this measure to be useful on the Level Diagrams, the base ten logarithm of this value is used.

C. Optimizer Setup

The optimizer was run for 20 generation with a crossover rate and mutation rate of 0.15.

The input parameters for the PI controller are the *P* and *I* as shown in (11).

$$K(s) = P + \frac{I}{s} \tag{11}$$

The parameters are bounded to the range $P, I \in [0,10]$.

Similarly for SMC, the input parameters are λ, Φ and *K*. They are bounded to the ranges $\lambda \in [0,4]$, $\Phi \in [0,6]$ and $K \in [0,10]$.

These limitations on λ and Φ are introduced based on the analysis of the eigen values of the resulting system dynamics. They are primarily introduced as a result of limited sampling speed.

The PI cancel set's parameters are specified such that on the range [0,10] pole – zero cancellation occurs for all controllers.

D. *ODE Solver Setup*

The ODE solver uses an RK4 algorithm with a sampling time of 0.05.

E. *Pole-Zero Cancel Set*

The pole-zero cancelling PI controller set is included as a reference set which is known to be undesirable in practice; since pole – zero cancellation can lead to poor internal performance [12], and as such should be avoided. Optimization does not take account of this and depends only on the costs being used. Pole-Zero cancellation typically results in very good output disturbance rejection behaviour but poor input disturbance rejection behaviour. Due to the good output disturbance rejection the controllers are often classified as optimal. Common costs used in Control Engineering optimization, such as the $ISE_{r,d}$ and $ISU_{r,d}$ costs used in this paper, do not account for this issue of internal performance. As such, optimizers will often optimize onto values where pole-zero cancellation occurs.

The costs included in this paper are unable to avoid this phenomenon, but they do show alternative solutions.

This is demonstrated below, where in Figure 4, two costs (Er and Ur) are used initially. The optimized PI controllers are shown in black, and the pole-zero cancellation set is shown in grey. From the figure one can see that all the optimal PI controllers lie on the pole-zero cancellation curve.

One thing to note is that the part of the curve where the optimal PI controllers does not lie is non-dominant and thus sub-optimal. It is included for visual completeness.

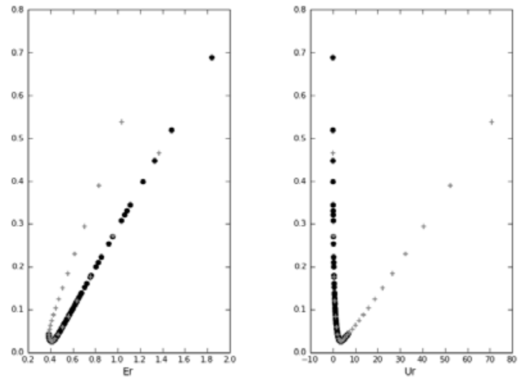


Figure 4: PI (black) and PI pole- zero cancellation (grey) cost Level Diagrams.

When two more costs (Ev and Uv) are included in Figure 5 the situation changes. Some of the resulting optimal controllers are still on the pole-zero cancellation curve (as they are still optimal), but additional optimal controllers are found not on this curve, thus avoiding the phenomenon.

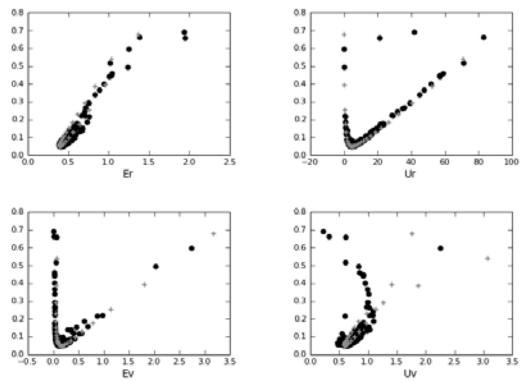


Figure 5: PI (black) and PI pole-zero cancellation (grey) cost Level Diagrams

IV. RESULTS

The results are split into two sections: General results from comparing the SMC and PI methods and results from using Level Diagrams.

A. *General Results*

During testing it was found that neither PI nor SMC dominated the other on any significant portion of their fronts.

This indicates that both Pareto Fronts exist in a mutually exclusive region of the n-dimensional space as shown in Figure 6.

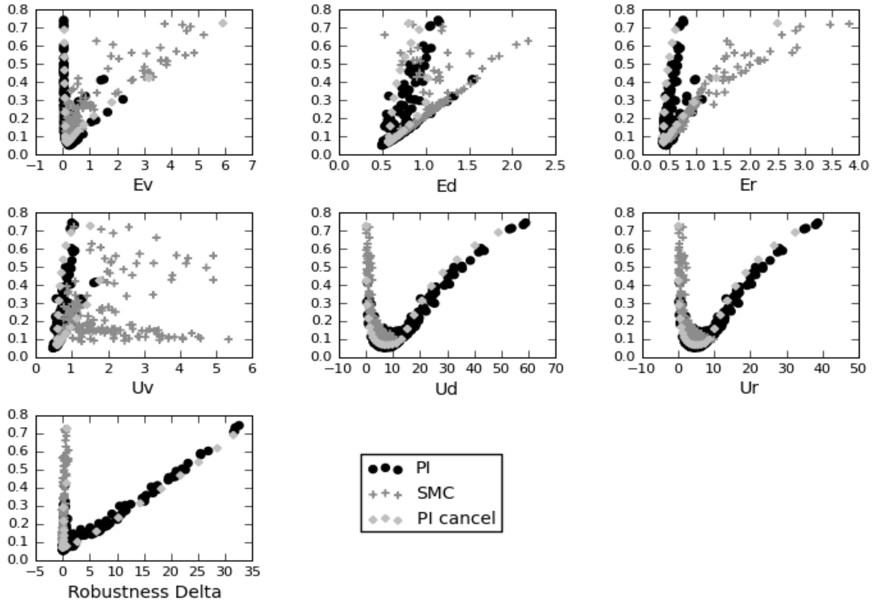


Figure 6: 2-D visualisation of separation between PI and SMC

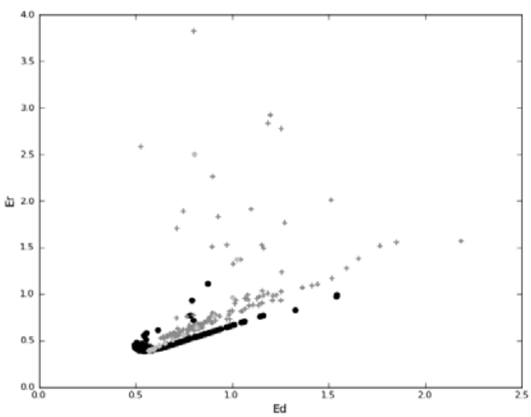


Figure 7: Cost Level Diagrams for PI, PI Cancel and SMC

B. *Level Diagram Results*

Figure 7 and Figure 8 shows the resulting Level Diagrams for the Pareto front for the PI (black), PI cancel (light grey) and SMC (dark grey) methods. In this case it is limited to a maximum Euclidean norm value of 0.75; this is done on the assumption that only reasonably balanced cases are to be examined as it effectively removes the extremes of the Pareto Fronts. The PI cancel set is retained as a known undesirable set.

A number of results can be drawn from visual inspection of the Level Diagrams.

Firstly, from the Robustness Delta Level Diagram, the SMC controller is much more robust to model changes than the PI controller; having a maximum value of less than 1, while a large portion of the PI solutions are very sensitive to model changes.

A number of observations can be drawn from the Level Diagrams for output disturbance and setpoint tracking (Ed, Ud, Er, and Ur). The PI controllers respond in a very similar fashion for both setpoint and disturbance; this is expected as there are no modelled feedback dynamics in the simulation. However, the SMC controller's ISE for the two disturbances changes: this is a function of this controller design.

Further, on the Ud and Ur Level Diagrams, SMC generally requires less controller effort *u*, having a maximum value of around 15 and 12 respectively. While a large number of the PI controller solutions require far larger controller effort to achieve their control.

That being said, the PI achieves better Output disturbance rejection and setpoint tracking, shown on Ed and Er with their higher controller effort.

This situation differs for the input disturbance (Ev and Uv). PI manages equal or better input disturbance rejection as shown in Ev, but generally requires less controller effort than the SMC controller to a achieve it, shown in Uv.

A couple of observations can be drawn from the Input Level Diagrams shown in Figure 8. Firstly, lower values of lambda (λ) are undesirable as they limit the reaching speed on the SMC and thus would increase the various ISE measures. Higher values of lambda tend to have lower norm values.

The PI pole-zero cancellation set exists as a subset of the optimal PI controllers. This implies that further costs are needed.

While there are fewer results that can be drawn from the input plots in this case, they can be useful for design decisions and trade-off analysis.

Finally, as the proportional gain Kp for the PI controller increases above around 5, the controller norm worsens.

V. Conclusions

A number of general conclusions can be drawn with regard to both the comparison of PI and SMC design methods, and the use of Level Diagrams.

As expected, SMC proved far more robust in general that PI, and was clearly shown using Level Diagrams.

Including more costs does not prevent the phenomenon of pole-zero cancellation; including a set of pole-zero cancelling controllers in the Level Diagrams does help with identifying controllers that avoid this phenomenon.

While Integrated SMC degrades the disturbance rejection of SMC, it still retains the other robust characteristics of SMC. As such, it is a useful method for chatter avoidance in situations where its inferior disturbance rejection is deemed acceptable.

Level Diagrams proved to be a useful tool for visualisation and visual analysis of optimal controller sets. As with other visualisation techniques, Level Diagrams would benefit from interactive features that cannot be easily shown in print.

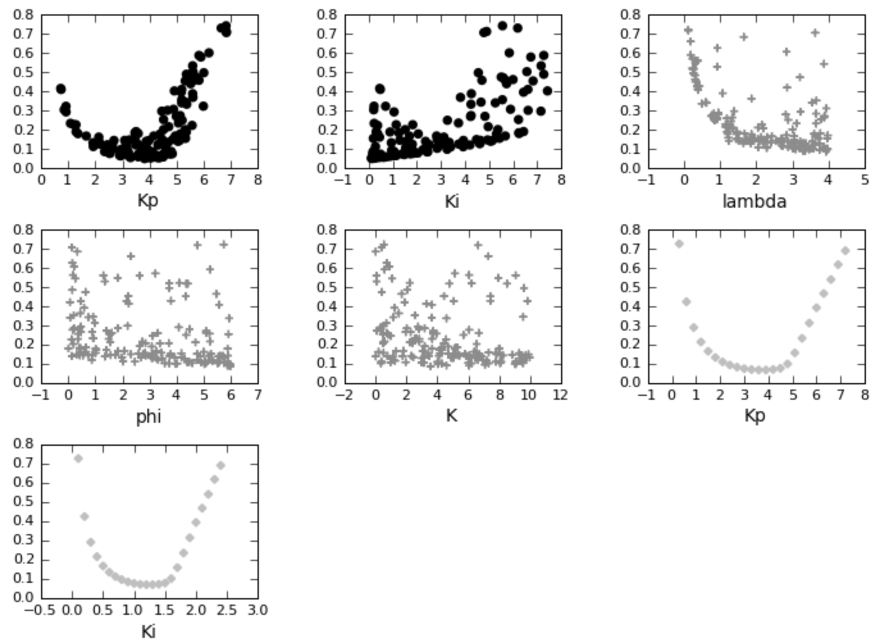


Figure 8: Input Level Diagrams for PI, PI Cancel and SMC

References

[1] M. Rapson, "Pareto Analysis of Controller Design Methodologies for Intergrator plus Dead Time Processes," *CISSE*, 2007.

[2] J.B. Yang, J.F. Whildborne G.P. Liu, *Multiobjective Optimization and Control*.: Research Studies Press LTD., 2002.

[3] J.M. Herrero, J. Sanchis, M. Marínez X. Blasco, "A new graphical visualization of n-dimensional Pareto front for decision-making in multiobjective optimization," *Information Sciences: an International Journal*, vol. 178, no. 20, pp. 3908-3924, October 2008.

[4] S.K. Spurgeon C. Edwards, *Sliding Mode Control: Thoery and Applications*.: CRC Press, 1998.

[5] Carlos A. Coello, "Twenty years of Evolutionary Multiobjective Optimization: A Historical View of the Field," *IEEE Computational Intelligence Magazine*, vol. 1, no. 2, pp. 28-36, 2006.

[6] Ruhul A. Sarker Hussein A. Abbass, "The Pareto Differential Evolution Algorithm," *International Journal on Artificial Intelligence Tools*, vol. 11, no. 4, pp. 531-552, 2002.

[7] Johan Andersson, "A Survey of Multiobjective Optimization in Engineering Design," 2000.

[8] Kenneth Price Rainer Storn, "Differential Evolution - A simple and efficient adaptive scheme for global optimization over continuous spaces," 1995.

[9] Cecil L. Smith, *Digital Computer Process Control*. Scranton: Intext Education Publishers, 1972.

[10] W. Li J.J.E. Slotine,.: Prentice Hall, 1991, ch. 7, pp. 276-309.

[11] V.I. Utkin, Ü. Özgüner K.D. Young, "A Control Engineer's Guide to Sliding Mode Control," *IEEE Trans on Control Systems Technology*, vol. 7, no. 3, pp. 328-342, 1999.

[12] P.A. La Plante, *Comprehensive dictionary of electrical engineering*.: CRC Press & IEEE Press, 1999.

[13] Keith J. Burnham Ali J. Koshkouei, "Control of DC Motors Using Proportional integral sliding mode,".

Interaction with Objects in Virtual Environments

Jesús A. Verduzco-Ramírez[1,] Enrique Araiza-Montes-de-Oca[2], Daniel A. Gaytán-Valencia[1], Nicandro Farías-Mendoza[1] and Bruno Raffin[3]

[1]Departamento de Sistemas y Computación, Instituto Tecnológico de Colima, Av. Tecnológico #1, C.P. 28976. Villa de Álvarez, Colima, México
[2]Facultad de Telemática, Universidad de Colima, Av. Universidad #333. Colonia Las Víboras. C.P. 28040. Colima, Colima, México
[3] INRIA. LIG. Avenue J. Kuntzmann. 38330 Montbonnot. France.

***Abstract*— The applications of virtual and augmented reality require high performance equipment normally associated with high costs, inaccessible for small organizations and educational institutions. By adding computing power and storage capacity of several PCs interconnected by a network we have an alternative of high performance and low cost. The objective of this project is to implement an augmented reality platform integrated with items available in the laboratories of an educational institution, such as PCs, webcams, Ethernet networking and free software. This platform is an attractive and economical alternative for the development and implementation of virtual and augmented reality applications to enable low-budget institutions simulate processes, physical phenomena, weather, etc. In this first phase, we have a prototype that creates a scene consisting of a set of virtual objects, the video stream acquired from a Webcam and processes the captured image into a graphic with a set of virtual objects. A user facing the angle of acquisition of the webcam can interact with virtual objects in the scene with only the move of his hand. Although our prototype is reduced to only two computers the first in which images are acquired through the webcam and the second where the mixture, processing and visualization of the scene is made, these early results demonstrate the feasibility of our idea and are a strong motivation for continuing the project. As future work we have planned to increase the number of Webcams, to explore the interaction with complex virtual scenes as well as distributing the processing of the scene among a group of computers to increase the size of the scene displayed.**

***Index Terms*—Augmented reality, virtual reality, interactive applications, computer graphics.**

I. Introduction

Virtual reality (VR) is a technology that superimposes virtual information on the user's vision. Potential applications have place in many fields such as medicine, military and entertainment.

The goal of VR is to create an experience that makes the user feel is immersed in a virtual world apparently real, for it, VR uses 3D graphics and sound that surrounds the scenes shown [1]. The VR uses the vision of the user, who moves into the virtual world using appropriate devices, such as goggles or electronic gloves. The VR exploits all the techniques of image reproduction and extends them, using them within the environment in which the user can examine, manipulate and interact with the objects shown [2].

In addition to the VR, there is the technology of augmented reality (AR), which consists of a collection of devices that enhance the capabilities of human senses [3]. The AR is a technology that integrates elements of the environment typically captured in video format and audio with three-dimensional computer graphics objects resulting in new consistent, supplemented and enriched worlds [4].

Both technologies are closely related but have differences. VR allows immersion of the user in an artificial world that completely replaces the user's real world. The RA maintains the user's real world enriched by the presence of virtual items. The applications of this technology are found in the following areas: military, medical, recreational, archaeological and pedagogical.

There are several platforms that use this approach successfully; one of the most outstanding is Grimage [5], which performs 3D modeling of images acquired by multiple cameras, physical simulation and parallel execution that offers a new immersive experience. On this platform it is possible to capture objects to be instantaneously modeled in 3D and entered into a virtual world populated by rigid or flexible solid objects with possible interaction with them. Grimage requires for their operation high-performance equipment. The existing configuration at the INRIA Rhone Alpes has six high definition cameras, a wall of images with 16 video projectors and two interconnected clusters, one with 11 dual-Xeon 2.6 GHz for the application and another with 16 dual-Opteron 2 GHz for the display [5]. Figure 1 shows one of the implementations of Grimage:

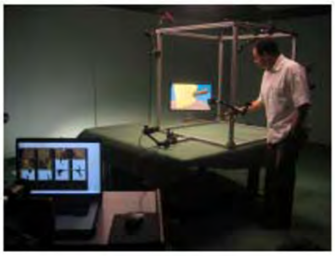


Fig. 1. Interacting with virtual objects in Grimage.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_37, © Springer Science+Business Media B.V. 2010

The software used in Grimage, FlowVR [6] is a VR API that offers support for heterogeneous codes executed in parallel. FlowVR allows the development and implementation of interactive applications in high performance computing clusters and grids. FlowVR reuses and extends the data flow paradigm commonly used for scientific visualization environments [7, 8].

One of the limitations of the AR platforms such as Grimage is that they are made up of costly and inaccessible equipment for educational institutions that have a small budget. Our approach aims to avoid such problems. Our proposal is to develop an augmented reality platform that integrates low-cost equipment such as Ethernet, PC, Webcams and free software. One use expected to achieve with the implementation of this platform is that teachers of an educational institution can simulate physical models and manipulate virtual objects in applications that are intended to support teaching.

II. ARCHITECTURE OF OUR SOLUTION

To implement a platform for low-cost RA, our solution is divided into four functional modules, as shown in Figure 2:

- One or more webcams to capture the video stream.
- A module for the acquisition and distribution of images captured with a webcam.
- A rendering module to mix the flow of video and interactive 3D scenes.
- A display wall for viewing the results.

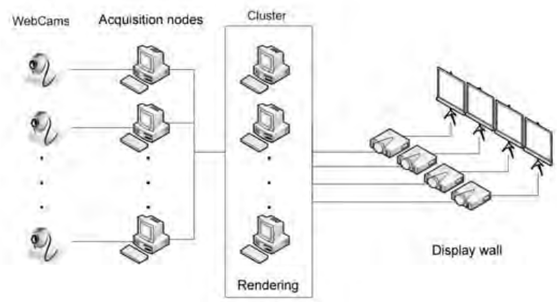


Fig. 2. Interacting with virtual objects in Grimage.

In this first phase, to verify the feasibility of our idea we have implemented a prototype consisting of only two nodes.

A. Acquisition Node

The acquisition node connects one or more web cameras, which acquire images of the environment. A video server stores and transmits these images to the rendering node.

B. Rendering Node

The rendering node is running an instance of FlowVR, it receives the video frames, makes the rendering and subsequent display through MPlayer immersed in FlowVR.

C. Operation

As described in Section A, the node captures images and sends them in a video stream to the rendering node.

As a video server we selected Spook [9], a free software that allows the delivery of video streams over IP networks. Spook takes the images stored in the buffer, compresses the data in sequential order, and sends them as RTP packets through the network. In the rendering node, the client starts the MPlayer immersed in FlowVR. Thus, it is possible to display the images captured by the webcam in the rendering the node.

The mixture of images captured with the webcam and a 3D scene component is provided by the FlowVR Render, which allows the combination of the video stream to a virtual environment rendered by FlowVR.

III. IMMERSION OF OBJECTS

A. Description of our Application

To test the functionality of our platform, we have developed some examples that demonstrate the recognition of the contours of the objects captured by the webcam and the interaction with virtual objects without using any device of interaction. For both operations we used the OpenCV library [10].

B. Contours detection

A contour is a list of points representing the curvature of an image. This representation may be different depending on the circumstances. There are many ways to represent them. In OpenCV the contours are represented by sequences where each entry in that sequence encodes information about the location of the next point on the curve.

To get to show the contours of an image, we first need a method for managing memory. OpenCV has an entity called Memory Storage, which is a method for managing the location of dynamic objects in memory. Any object of OpenCV that requires the use of memory as part of its normal operation, needs a Memory Storage. The Memory Storage is a linked list of memory blocks for fast data storage and retrieval.

The type of object that can be stored in a memory storage is a sequence. The sequences are linked lists of other data structures.

The first developed example is the recognition of a series of letters and figures drawn by hand. Examples are shown in Figure 3.

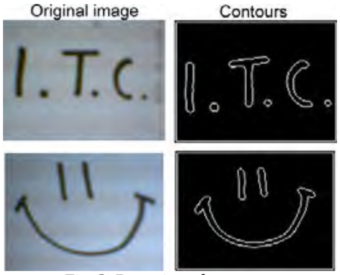


Fig. 3. Detection of contours.

C. *Interacting with Virtual Objects*

By using our prototype it is possible to capture objects using the webcam and embed them into a 3D application, which provides opportunities to interact with virtual objects that make up the 3D scene. In an application of this type we identified three elements, shown in figure 4:

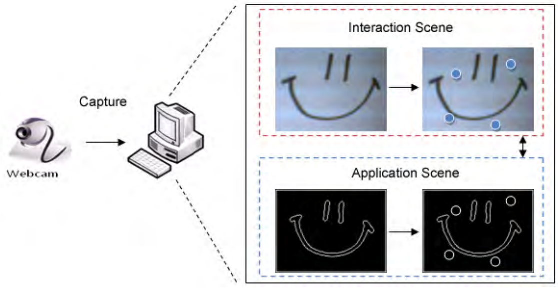


Fig. 4. Architecture of the applications developed.

• The interaction scene: represented by a set of images from the camera.
• The application scene: represented by a set of images in which there are some features of the scene of interaction. For example, contours, corners and colors.
• Finally, the virtual objects embedded in scenes of interaction and application, which react to events within the application stage and are represented through images or textures in the scene of interaction.

To test the interaction with virtual objects, we developed a couple of examples intended to show how a user can interact with them.

The first application shows a set of bubbles that the user can exploit to make moves on the interaction scene. The application scene is the difference between a frame image and its predecessor. Thus, a white shadow is formed as shown in Figure 5.

The second application example shows a set of balls bouncing in the interaction scene. When the user touches one or when they collide with each other, they change their direction.

The application scene consists of the contours of the figures appearing in the interaction scene, which upon contact with virtual objects, they are running a routine to change its direction of motion.

Both example applications are shown in figure 5.

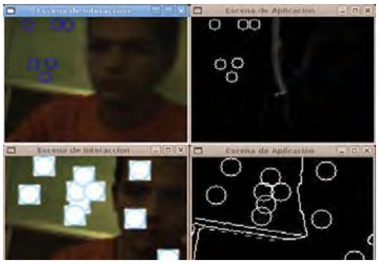


Fig. 5. Example applications, to the left are the interaction scenes and to the right the application scenes.

IV. Conclusion

In this paper we describe the first phase in the development of an augmented reality platform of low cost.

Our solution uses low cost components, present in the laboratories of almost any educational institution, such as, data networking, Webcams, PCs and free software, which makes our proposal an attractive and relatively simple to implement.

This platform allows educational institutions; simulate processes, physical phenomena, weather, etc.
A user in front of a webcam can interact with virtual objects in an application without using any interaction device.

V. Future Work

The results obtained so far encourage us to continue this project. We want to increase the number of nodes involved in the rendering scenes, the number of cameras allowing the acquisition of more than one video stream, to add textures to objects displayed on the interaction scene and study the system performance in these applications.

References

[1] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart. The Cave Audio Visual Experience Automatic Virtual Environment. Communication of the ACM, 35(6):64_72, 1992.

[2] J. Hilera, S. Otón, J. Martínez, “Aplicación de la Realidad Virtual en la Enseñanza a Través de Internet”, Universidad de Alcalá, septiembre de 1999. [Online]. Disponible en: *http://www.ucm.es/info/multidoc/multidoc/revista/num8/hilera-oton.html* [Accesado en diciembre 10, 2008].

[3] J. Ahrens, C. Law, W. Schroeder, K. Martin, and Michael Papka, A Parallel Approach for Efficiently Visualizing Extremely Large, Time-Varying Datasets. http://www.acl.lanl.gov/Viz/papers/pvtk/pvtkpreprint/.

[4] L. Heras, J. Villareal, “La Realidad Aumentada: Una Tecnología en Espera de Usuarios”, Revista Digital Universitaria, vol. 5, no. 7, Agosto, 2008.

[5] J. Allard, C. Menier, B. Raffin, E. Boyer, F. Faure. 2007. Grimage: Markerless 3D Interactions. ACM SIGGRAPH 2007 emerging technologies.

[6] J. Allard, V. Gouranton, L. Lecointre, S. Limet, E. Melin, B. Raffin and S. Robert. 2004. FlowVR: a Middleware for Large Scale Virtual Reallity Applications. In Euro-Par 2004 Parallel Processing: 10th International Euro-Par Conference, 497-505. http://flowvr.sf.net.

[7] Covise Programming Guide, 2001. http://www.hlrs.de/organization/vis/covise.

[8] S. Jordà. Realidad Aumentada. Taller de Sistemes Interactius 2003 – 2004. Universitat Pompeu Fabra.

[9] Spook Video Live Streamer http://www.litech.org/spook/ “OpenCV Wiki”, IBM, 2002. [Online]. Disponible en: *http://opencv.willowgarage.com/wiki/* [Accesado en mayo 11, 2009]

One Approach for Training of Recurrent Neural Network Model of IIR Digital Filter

Stefanova S.A.

Department of Electronic Engineering and Technologies, Technical University of Sofia,
8 Kliment Ohridski Str.,
Sofia, 1797, Bulgaria

***Abstract* - One approach for training of recurrent neural network model of 1-D IIR digital filter is proposed. The sensitivity coefficients method has been applied in the training process of the neural network. The set of time domain data is generated and used as a target function in the training procedure. The modeling results have been obtained for two different cases - for 4-th order bandpass IIR digital filter and for partial response IIR digital filter. The frequency domain behavior of the neural network model and the target IIR filter has been investigated. The analysis of the frequency responses shows good approximation results.**

I. INTRODUCTION

In the last years various approaches for modeling and simulation of the analog and discrete dynamic systems based on different type of neural networks have been developed. Basic results related to the discrete dynamical systems approximation using neural networks are discussed in [1]. Many approaches based on neural network have been successfully applied to the problems of design, modeling and simulation of analog filters [2], linear and non-linear [3], [4] digital filters,1-D and 2-D non-recursive (FIR) digital filters [5], [6], [7], [8], 1-D and 2-D recursive (IIR) digital filters [9], [10], [11], [12] and adaptive digital filters [4].

One approach for the 1-D FIR digital filter design based on the weighted mean square method and neural network to state the approximation problem is proposed in [5]. The optimal design approaches of high order FIR digital filters based on the parallel algorithm of neural networks, which its activation matrix is produced by cosine and sine basis functions are given in [6],[7] . The main idea is to minimize the sum of the square errors between the amplitude response of the desired FIR filter and that of the designed by training the weight vector of neural networks, then obtaining the impulse response of FIR digital filter. The reference [8] provides an approach for the design of 2-D FIR linear-phase digital filters based on a parallel back-propagation neural networks algorithm.

One method for time domain design of 1-D and 2-D recursive digital filter, based on recurrent neural network is proposed in [9]. An approach for the training algorithm of a fully connected recurrent neural network where each neuron is modeled by an IIR filter is proposed in [10]. The weights of each layer in the network are updated by optimizing IIR filter coefficients and the optimization is based on the recursive least squares method. A Hopfield-type neural network for the design of IIR digital filters with given amplitude and phase responses is discussed in [11]. A method for 2-D recursive digital filters design is investigated in [12], where the design problem is reduced to a constrained minimization problem the solution of which is achieved by the convergence of an appropriate neural network.

Some methods for the non-linear digital filters design using neural networks are considered in [3]. A design method for nonlinear adaptive digital filters using parallel neural networks is given in [4].

A various methods have been developed for the purposes of the training process of the recurrent neural networks [13], [14], [15]. The back-propagation through time method is the basic approach for recurrent neural network training, discussed in [13]. Some training algorithms based on the sensitivity theory are considered in [14], [15]. The application of the Lagrange multipliers method as a training procedure of the recurrent neural network model of IIR digital filter is discussed in [14], [16]. Reference [9] describes the learning algorithm for the neural network structure based on the state space representation of 1-D and 2-D IIR digital filters.

In this paper one approach for training of IIR digital filter model based on two layer recurrent neural network is proposed. The structure and mathematical description of the recurrent neural network is given in section II. The training process of the neural network using sensitivity coefficients method is discussed in Section III. The neural network model has been trained in such a way that with given predetermined input signal, the output variable approximates the target function in mean square sense. The neural network model of the IIR digital filter is considered in section IV. The effectiveness of the proposed neural network model of the IIR digital filter is demonstrated in Section V. The modeling results have been obtained for two different cases - for 4-th order bandpass IIR digital filter and for partial response IIR digital filter which is a special case of the Nyquist recursive digital filter [17]. The frequency domain behavior of the neural network model and the target IIR filter has been investigated and the analysis of the frequency responses shows good approximation results. Some simulations are realized using harmonic signals with different frequencies in the filter's passband and stopband. The proposed approach can be extended and successfully applied to the case of modeling of nonlinear IIR digital filters.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_38, © Springer Science+Business Media B.V. 2010

II. Structure and Description of the Neural Network

The two layer recurrent neural network is considered. The structure of neural network is shown in Fig. 1.

The recurrent neural network is described with the recurrent system of equations as follows:

$$\begin{bmatrix} s_1(k) \\ s_2(k) \\ \cdots\cdots \\ s_n(k) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \ldots & w_{1n} \vdots \ldots w_{1n_z} \\ w_{21} & w_{22} \ldots & w_{2n} \vdots \ldots w_{2n_z} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \vdots \cdots\cdots \\ w_{n1} & w_{n2} \ldots & w_{nn} \vdots \ldots w_{nn_z} \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \cdots\cdots \\ x_n(k) \\ u_1(k) \\ \cdots\cdots \\ u_{n_u}(k) \end{bmatrix}, \tag{1}$$

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \cdots\cdots \\ x_n(k+1) \end{bmatrix} = \begin{bmatrix} f(s_1(k)) \\ f(s_2(k)) \\ \cdots\cdots \\ f(s_n(k)) \end{bmatrix} \quad k = k_0, k_0+1, \ldots, k_f - 1,$$

where f(.) is the activation function for the first layer of the neural network

Let be introduced the vector function **f(s)** as follows:

$$\mathbf{f(s)} = [f(s_1), f(s_2), \ldots, f(s_n)]^T,$$

then the neural network description in the matrix form is:

$$\mathbf{x}(k+1) = \mathbf{f}(\mathbf{Wz}(k)), \quad \mathbf{x}(k_0) = \mathbf{x}_0, \tag{2}$$

$$\mathbf{y}(k) = \mathbf{Cx}(k), \tag{3}$$

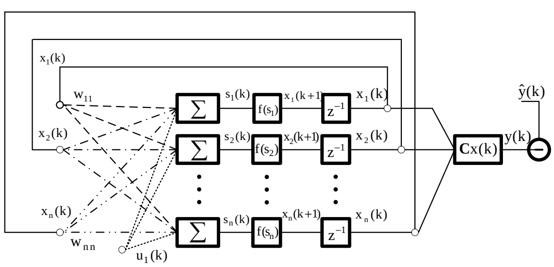


Fig. 1. Recurrent neural network structure for digital filter modeling

where

$\mathbf{x} \in \mathbf{R}^n$ is a state space vector,

$\mathbf{u} \in \mathbf{R}^{n_u}$ is a vector of neural network input excitations,

$\mathbf{y} \in \mathbf{R}^m$ is a vector of neural network output,

$\mathbf{W} \in \mathbf{R}^{n \times n_z}$ and $\mathbf{C} \in \mathbf{R}^{m \times n}$, $m \le n$, $\mathbf{C} = \lfloor c_{ij} \rfloor$,

$c_{ij} = \begin{cases} c_i & i = j \\ 0 & i \ne j \end{cases}$ are matrixes of system coefficients,

n - is a number of neurons at the first layer,

n_u - is a number of input excitations,

$n_z = n + n_u$ - is a number of neural network inputs.

The equations (2), (3) can be used to describe the recurrent neural network shown in Fig. 1.

The following additional vectors have been defined:

- a vector of the first layer inputs of neural network

$$\mathbf{z}(k) = [x_1(k), x_2(k), \ldots x_n(k), u_1(k), u_2(k), \ldots, u_{n_u}(k)]^T,$$

$$\mathbf{z} \in \mathbf{R}^{n_z} \tag{4}$$

- a vector of the neural network outputs

$$\mathbf{y}(k) = [y_1(k), y_2(k) \ldots y_m(k)]^T \tag{5}$$

- a vector of the neural network weighting coefficients

$$\mathbf{p} = \lfloor w_{11}\ w_{12} \ldots w_{1n_z}, w_{21}\ w_{22} \ldots w_{2n_z}, \ldots w_{nn_z}, c_1, c_2 \ldots c_m \rfloor,$$

$$\mathbf{p} \in \mathbf{R}^{n_p} \tag{6}$$

where $n_p = n.n_z + m$ is the number of neural network coefficients and the elements of matrix $\mathbf{W} \in \mathbf{R}^{n \times n_z}$ are introduced row by row.

Mean square error objective function is defined in following form:

$$J(\mathbf{p}) = \frac{1}{2} \sum_{k=k_0}^{k_f-1} \sum_{i=1}^{m} [y_i(k) - \hat{y}_i(k)]^2, \tag{7}$$

where $\{\hat{y}_i(k)\}$ is a set of samples of the target .

III. Algorithm of Sensitivity Coefficients

The main problem in the neural network training process is the gradient calculation of the mean square objective function (7) with respect to weights.

For that reason, it is necessary to determine the gradient components of this objective function.

$$\frac{\partial J(p)}{\partial p_l} = \sum_{k=k_0}^{k_f-1} \sum_{i=1}^{m} [y_i(k) - \hat{y}_i(k)] \frac{\partial y_i}{\partial p_l}, \quad l=1,\ldots,n_p \tag{8}$$

For the first layer of the neural network the derivatives $\frac{\partial y_i}{\partial p_l}$ have been determined from (9) as follows:

$$\frac{\partial y_i(k)}{\partial p_l} = c_i \frac{\partial x_i(k)}{\partial p_l}, \; i = 1, \cdots, m, \; l = 1, \cdots n_w, \; n_w = n.n_z \tag{9}$$

The derivatives $\frac{\partial x_i(k)}{\partial p_l}$ from (9) can be obtained from the recurrent system:

$$\frac{\partial x_i(k+1)}{\partial p_l} = \frac{\partial x_i(k+1)^f}{\partial p_l} + \sum_{j=1}^{n} \frac{\partial x_i(k+1)^f}{\partial x_j(k)} \frac{\partial x_j(k)}{\partial p_l};$$

$$; i = 1, \ldots, n; \; l = 1, \ldots, n_w; \tag{10}$$

$$\frac{\partial x_i(k_0)}{\partial p_l} = 0 \quad ; \quad k = k_0, k_0+1, \ldots, k_f - 1$$

where

$$\frac{\partial x_i(k+1)^f}{\partial w_{qj}} = \begin{cases} z_j . f'(s_i)\,; & i = q = 1, \cdots, n\,; \; j = 1, \cdots, n_z \\ 0 \quad ; & i \ne q\,; \quad j = 1, \cdots, n_z \end{cases} \tag{11}$$

$$\frac{\partial x_i(k+1)^f}{\partial x_j(k)} = w_{ij} f'(s_i)\,; \quad i = 1, \cdots, n\,; \quad j = 1, \cdots, n \tag{12}$$

and the derivatives (11), (12) are determined from the neural network shown in Fig. 1 in the case when network structure has been considered as a feedforward neural network.

For the second layer of the neural network the derivatives $\frac{\partial y_i}{\partial p_l}$ have been determined as follows:

$$\frac{\partial y_1(k)}{\partial p_{n_w+1}} = x_1(k)\,, \cdots, \frac{\partial y_m(k)}{\partial p_{n_p}} = x_m(k) \qquad (13)$$

Then the objective function (7) can be minimized applying the standard optimization procedure.

There are two ways to realize the training algorithm of the neural network. The first one, described above, insists on determining the solution of the recurrent system (1) and calculating the summary error from (7) for all time samples $\mathrm{k} = \mathrm{k}_0, \mathrm{k}_0 + 1, \ldots$. Simultaneously with this calculation process, the sensitivity coefficients from (10) have to be calculated and the gradient components of the objective function (7) have to be obtained using equations (8). After that the objective function (7) has to be minimized applying standard gradient procedure. This training procedure supposes the processing of all calculated data i.e. the training process is off-line. The second way to train the neural network requires calculation of the error function (7) on the neural network output by solving the system (1) for every time sample $\mathrm{k} = \mathrm{k}_0, \mathrm{k}_0 + 1, \ldots$. The gradient (8) of this error function has to be determined using the sensitivity coefficients obtained from system (10). The next operation is a correction of the neural network weights using the components of the error gradient (8). This training process is recurrent or on-line training which can be realized in the real time. The possibility of on-line realization of the training procedure is the main advantage of the sensitivity coefficients algorithm in comparison with the Lagrange multipliers training method, described in [16], when the training process can be realized only off-line.

IV. Neural Network Model of IIR Digital Filter

The IIR digital filter can be considered as a linear discrete dynamic system described in time domain with n – order difference equation that has been stated using the delayed samples of the input excitation and the response signal at the output. Transforming the difference equation the state space description of the IIR digital filter can be obtained as a linear system of “n” number of first order recurrent equations. The advantages of the state space representation are the simplicity of time domain analysis, the matrix form description of digital filter impulse and step responses and possibility of parallel calculations.

The neural network structure shown in Fig. 1 must be modified to linear structure. In this case the activation function f(x) is linear. The corresponding system of equations (1)-(3) is linear recurrent system. The sensitivity coefficients algorithm modified for linear network is used as a training procedure of the neural network model.

V. Modeling Results

Case 1. Modeling of 4 th order bandpass IIR digital filter

The following requirements to the magnitude response of the target bandpass digital filter are specified:
passband [300–3400] Hz, passband loss – 3 dB; stopband [0–100] Hz and [3600–4000] Hz, stopband loss – 15 dB; sampling frequency – 8000 Hz.

The training sequences are generated using the impulse response of target IIR digital filter shown in Fig. 2. This target IIR digital filter has been designed by MATLAB’s Filter Design Toolbox.

The model of the bandpass IIR filter is realized as recurrent neural network structure with 4 neurons.

The effectiveness of the proposed algorithm is demonstrated by simulation with neural network model. The simulation is realized applying input harmonic excitations with frequencies in the digital filter’s passband and stopband. These sinusoidal signals are different from the set of data used in the training process of neural network model. The impulse responses of the target digital filter and the neural network model are shown in Fig. 2 and Fig. 3.

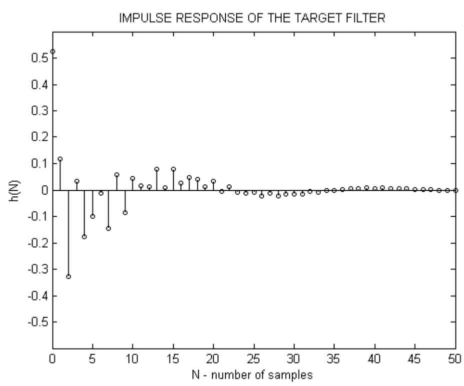


Fig 2. Impulse response of the target bandpass IIR filter

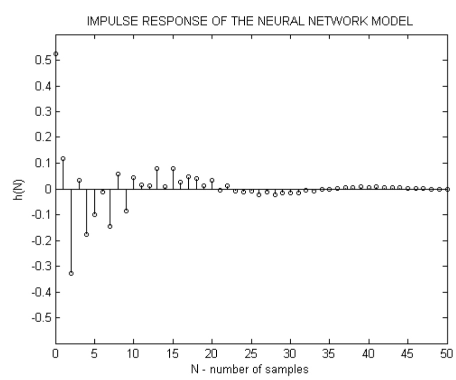


Fig. 3. Impulse response of the neural network model

The magnitude responses in frequency domain of the target IIR bandpass digital filter and neural network model are shown in Fig. 4 and Fig. 5, respectively.

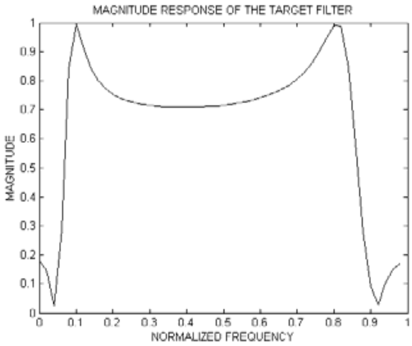


Fig 4. Magnitude response of the bandpass target filter

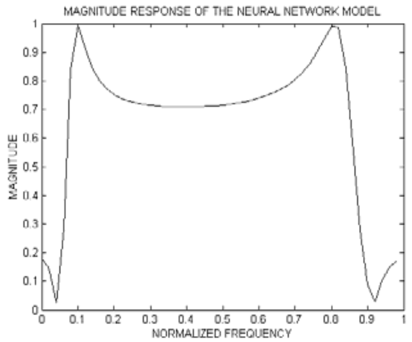


Fig. 5. Magnitude response of the neural network model

Some simulation experiments have been implemented when on the input of neural network model have been applied sinusoidal signals with different frequencies and normal distributed white noise N (0, 0.1) added to the harmonic excitation. The input harmonic signal with frequency 60 Hz chosen from the digital filter's stopband and output response of the neural network model are shown in Fig. 6.

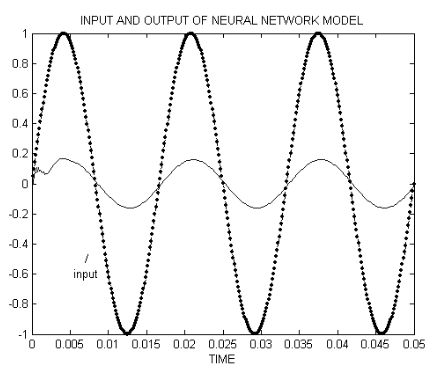


Fig. 6. Input signal with frequency 60 Hz and output response of neural network model

The comparison between the input harmonic excitation with frequency 700Hz chosen from digital filter's passband and the output reaction of neural network model is illustrated in Fig. 7.

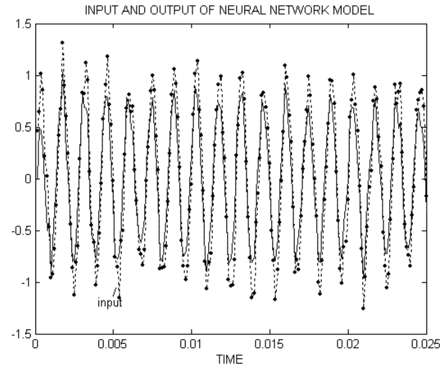


Fig 7. Input signal with added noise and frequency 700 Hz and output response of neural network model

The time responses of the target bandpass IIR digital filter and the neural network model in the case of input signals with frequencies 50 Hz and 1kHz are given in Fig. 8 and Fig. 9.

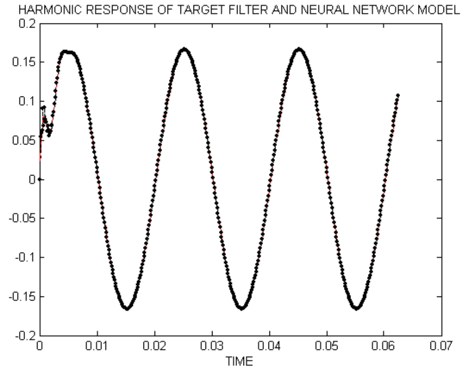


Fig. 8. Harmonic responses of target filter and neural network model – 50 Hz

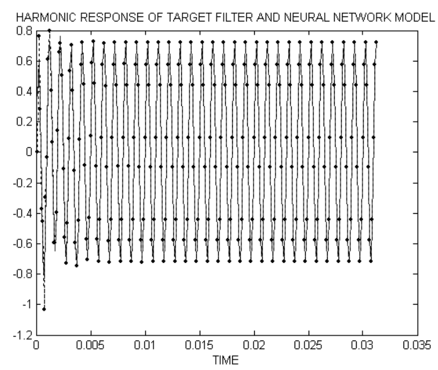


Fig. 9. Harmonic responses of target filter and neural network model – 1 kHz

Case 2. Modeling of partial response IIR digital filter

The class 1 partial response IIR digital filter is the special case of the IIR Nyquist filter [17].

Nyquist recursive digital filter plays an important role in digital data transmission for its intersymbol interference (ISI)-free property. Also it can be adopted in decimation or interpolation multirate systems. To achieve zero ISI, Nyquist filter must satisfies some criteria in time domain that they should have zeros equally spaced in the impulse response coefficients except one specified.

Impulse response $h(n)$ of the (IIR) Nyquist filter with the time domain constraints is defined in the form [17]:

$$h(K+kN)=\begin{cases}\dfrac{1}{N}\neq 0\,, & if\;\; k=0\\ 0\,, & otherwise\end{cases} \tag{14}$$

where K and N are integers

The transfer function $H(z)$ of the (IIR) Nyquist filter can be expressed as:

$$H(z)=b_K z^{-K}+\frac{\sum\limits_{i=0,\,i\neq kN=K}^{N_n} b_i z^{-i}}{\sum\limits_{i=0}^{N_d/N} a_{iN} z^{-iN}}\,,\quad b_K=\frac{1}{N} \tag{15}$$

where N_n, N_d are integers, all the filter coefficients a_i, b_i are real, $a_0=1$, N_d is the multiple of N.

The specifications for the IIR partial response filter are the following: $N_n=10, N_d=2,\ N=2, K=6$ [17].

The impulse response of the partial response filter, shown in Fig. 10, is used as target function in the training procedure of the neural network model. The magnitude response of the partial response IIR digital filter has a cosine shape which is illustrated in Fig. 11.

The impulse response and magnitude response of the neural network model with 6 neurons of IIR partial response filter are shown in Fig.12 and Fig. 13 respectively.

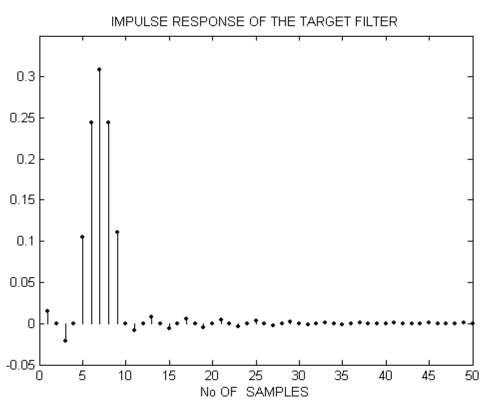


Fig 10. Impulse response of the target IIR partial response filter

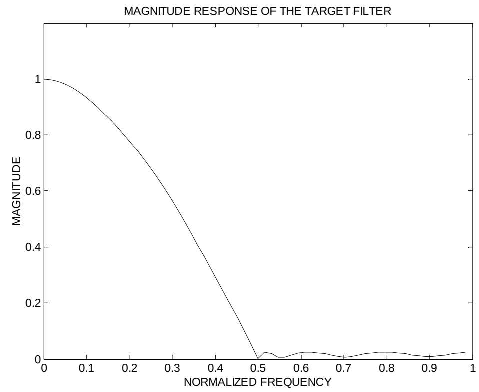


Fig 11. Magnitude response of the target IIR partial response filter

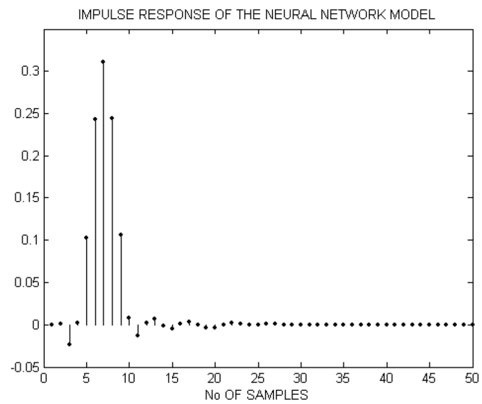


Fig 12. Impulse response of the neural network model of IIR partial response filter

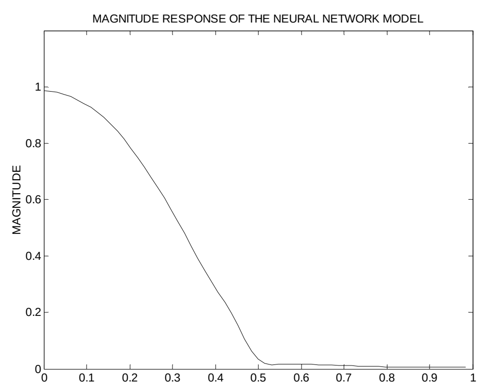


Fig.13. Magnitude response of the neural network model of IIR partial response filter

VI. Conclusion

One approach for the IIR digital filter modeling based on recurrent neural network is proposed. The sensitivity coefficients method has been used to develop an effective training algorithm for the neural network model. The training process also can be on-line and has a real time realization. The partial response IIR and bandpass IIR digital filters have been modeling using the recurrent neural network structures. To investigate the behavior of the digital filter models some simulations have been realized using input harmonic excitations with different frequencies. The analysis of the obtained results shows a very good approximation of the impulse responses in time domain and the magnitude responses in the frequency domain. This approach can be extended and successfully applied to the case of modeling of nonlinear recursive digital filters.

References

[1] Jin L., P. Nikiforuk, and M. Gupta, "Approximation of discrete-time state-space trajectories using dynamical recurrent neural networks", *IEEE Transactions on Automatic Control,* Vol.40, No 7, 1995.

[2] H. Cruse , *Neural Networks as Cybernetic Systems*, Brains, Minds and Media, Bielefeld, Germany, October 2006.

[3] M. Pedersen, "Optimization of recurrent neural networks for time domain series modeling", *Ph.D Thesis, Department of Mathematical Modeling*, Technical University of Denmark, 1997.

[4] J. Cao, Yahagi T., "Nonlinear adaptive digital filters using parallel neural networks", *Proceedings of IEEE International Conference on Neural Networks*, Vol. 2, Nov-Dec 1995 pp. 850 – 853.

[5] D. Bhattacharya, and A. Antoniou, Real-time design of FIR filter by feedback neural network, *IEEE Signal Processing Letters*, Vol.3, No 4, May 1996, pp. 158 – 161.

[6] Zhe-Zhao Zeng; Ye Chen; Yao-Nan Wang, "Optimal Design Study of High-Order FIR Digital Filters Based on Neural-Network Algorithm", *International Conference on Machine Learning and Cybernetics*, 13-16 Aug. 2006, pp 3157 – 3161.

[7] Zeng Zhe-zhao; Wen Hui, "Optimal design study of three-type FIR high-order digital filters based on sine basis functions neural-network algorithm", *IEEE International Symposium on Communications and Information Technology*, Vol. 2, No 12-14 ,pp. 921- 924, October 2005.

[8] X. H. Wang, Y. G. He, H. Li, Y. L. Peng, M.J. Li, "Design of 2-D digital filters using back propagation neural networks", *Proceedings of International Conference on Machine Learning and Cybernetic,* 18-21 Aug. 2005, Vol. 8, pp 4684 – 4689.

[9] Wang D. and A. Zilouchian, *Intelligent control systems using soft computing methodologies*, Editors: Ali Zilouchian, Mo Jamshidi**,** Prentice Hall, March 2001, pp, 93-110.

[10] Chow, T.W.S. Siu-Yeung Cho "An accelerated recurrent network training algorithm using IIR filter model and recursive least squares method" *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications,* Vol. 44, No 11, Nov, 1997, pp. 1082-1086.

[11] Bhattacharya, D.; Antoniou, A., "Design of IIR filters with arbitrary amplitude and phase responses by feedback neural networks", *IEEE International Symposium on Circuits and Systems, ISCAS '96*, Vol. 3, 12-15 May 1996, pp. 457 – 460.

[12] Mladenov, V.M., Mastorakis, N.E., "Design of two-dimensional recursive filters by using neural networks", *IEEE Transactions on Neural Networks,* Vol. 12, No 3, May 2001, pp. 585 – 590.

[13] W. T. Miller, R. S. Sutton and P. J. Werbos, *Neural Networks for Control*., Cambridge, MA:MIT Press, 1990.

[14] R. Grino, G. Cembrano and C. Torras, "Nonlinear system identification using additive dynamic neural networks – two on-line approaches", *IEEE Trans. on Circuits and Systems I*, vol. 47, no.2, 2000, pp. 150-165.

[15] Y. Fang, M. Yagoub, F. Wang and Qi-J. Zhang, "A new macromodeling approach for nonlinear microwave circuits based on recurrent neural networks", *IEEE Trans. on Microwave Theory and Techniques*, Vol. 48, no. 2, Dec. 2000, pp. 2335-2344

[16] Stefanova S., "One Dimensional IIR digital filter modeling based on recurrent neural network", International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 08), University of Bridgeport, Bridgeport, U.S.A, December 5 - 13, 2008.

[17] S. Pei, Ch. Hsu, P. Wang, "Design of a class of IIR eigenfilters with time and frequency domain constraint, *IEEE Trans. on Circuits and Systems-II Analog and Digital Signal Processing*, vol. 49, No 2, February 2002, pp 145 - 151.

Classification of the Power Transformers using Dissolved Gas Analysis

T. C. B. N. Assunção, J. T. Assunção and A. D. Neto
Department of Electric Engineering, Federal University of São João del-Rei (UFSJ), Brazil
(e-mail: bessa@ufsj.edu.br, tarcisio@ufsj.edu.br, armandodiniz@gmail.com).

Abstract— Least squares support vector machines (LS-SVM), artificial neural networks (ANN) and the traditional methods based in the dissolved gas analysis were employed for the detection of incipient faults in power transformers. The diagnosis criteria's commonly used for dissolved gas analysis (DGA) in transformer insulating oil are Doernenburg Ratio Method, Rogers Ratio Method, IEC 60599, IEEE C57.104-1991 and Duval Triangle. In this paper, a comparative study using ANN, LS-SVM and DGA for the incipient fault detection in power transformers is presented. The artificial neural network and least squares support vector machines approaches results are compared with the DGA results obtained from analysis of the traditional methods already consolidated in the technical literature.

I. INTRODUCTION

Power transformers are high cost important equipment used in the transmission and distribution of the electric energy. Its right performance is important for the electric systems operation, because the loss of a critical unit can generate great impact in safety, reliability and in the cost of the electric energy supply, mainly under an ever-increasing competitive environment. The power transformers due to natural ageing, loading regime or electrical and mechanic efforts, operate under harder conditions than they were installed.

System power reliability improves when incipient faults in power transformers are identified and eliminated before they deteriorate to a critical operation condition.

Incipient transformer faults lead to electrical and thermal stresses on insulating materials. As a result of these stresses the insulating materials can breakdown and several gases are released on oil insulating. The analysis of these gases provides useful information about the fault conditions and the type of materials involved.

There are several techniques used in the detection those gases, but the dissolved gas analysis (DGA) was recognized as the most informative method. So, the DGA is the well-known diagnostic technique in the world because it's sampling and analyzing procedures are simple, inexpensive and easy to be standardized.

The analysis of specific types of gases in the oil-filled transformers is useful for early failure detection. The most important aspect of fault analysis is taking data that has been generated and correctly diagnosing the fault that is generating of gases.

To facilitate the diagnosis of incipient faults by DGA several methods of interpretation have been developed, as Rogers, Doernenburg, Key Gas, Duval Triangle, IEEE C57.104-1991, and IEC 60599 [1]. These methods are used for to interpret the concentration of gases and/or the relation between them, which are compared to thresholds to conclude about the state of transformer insulation. The gases analyzed are: Hydrogen (H_2); Methane (CH_4); Ethane (C_2H_6); Ethylene (C_2H_4); Acetylene (C_2H_2); Carbon Monoxide (CO) and Carbon Dioxide (CO_2).

The interpretation of the DGA is made by ratio methods that are coding systems that assign a certain combination of codes to a fault specific type. The codes are generated by calculating gas ratios and comparing the ratios to pre-defined ratio intervals. A fault condition is detected when a code combination fits the code pattern of the fault.

The entire phenomenon related at gases and their correlations with incipient faults have been characterized for imprecision, in doubt measurement and unmodelled non-linearity. So, there is necessity of an expertise to interpret the difficult or inconclusive DGA tests results.

In order to overcome this difficulty in the interpretation of tests results, there has been a substantial effort in the developing intelligent diagnostic in this area. This intelligent diagnostic is a combination of the artificial neural networks (ANN) and DGA samples for an incipient faults efficient detection in power transformers.

For this purpose ANN can be used, due to its learning capacity in the modeling complex and nonlinear relations [2]. ANN is submitted to a training process from real cases, and then handling appropriately new supplied data. Several ANN structures have been proposed by researchers that can be classified as static (SNN), dynamic temporal processing (TPNN), recurrent (RNN) and radial basis function (RBFN). The most well-know ANN configuration is the multi-layer feedforward network that have been applied successfully to solve some difficult and assorted problems including nonlinear system identification and control, financial market analysis, signal modeling, power load forecasting etc. The fast learning procedure and its great generalization capability have promoted the use of the ANN in the areas of non-linear identification, approximation and interpolation theory.

Recently, the Support Vector Machine (SVM) has been proposed as a new and promising technique for classification and regression of the linear and nonlinear systems [3].

The LS-SVM is a learning machine proposed in [3] corresponding a modified version of the SVM. Like the SVM, LS-SVM can be used in classification problems and

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_39,

approximation functions. The main characteristic of LS-SVM is smallest computational cost in relation to the ANN, without loss in the quality of the solutions.

In this paper, a comparative study of ANN, LS-SVM and DGA for the calculation of efficiency for the incipient fault detection is presented. The results of this study are useful in the development of a reliable transformer diagnostic system using intelligent systems and DGA.

II. INCIPIENTS FAULTS DIAGNOSTIC

The power transformers in operation are subject to failures due to electric and thermal stresses and the stresses due to environmental influences. The effect of any stress results in the ageing of insulation, it leads up to decomposition of the insulating oil and paper, with the generation of dissolved gases in the insulation oil.

The oil dissolved gases can signify a possible failure in the power transformers [1- 2].

The gases more generated inside power transformers during operation normal or during the degradation of isolation system are: Hydrogen (H_2); Methane (CH_4); Ethane (C_2H_6); Ethylene (C_2H_4); Acetylene (C_2H_2); Carbon Monoxide (CO) and Carbon Dioxide (CO_2); and they can indicate:

- Hydrogen (H_2): Corona;
- Hydrogen (H_2); Methane (CH_4); Ethane (C_2H_6); Ethylene (C_2H_4): Oil Thermal Decomposition;
- Carbon Monoxide (CO) and Carbon Dioxide (CO_2: Paper Thermal Decomposition;
- Acetylene (C_2H_2): Electric Arc.

In this paper were studied the methods more used for the dissolved gas analysis, as follows [3-5].

Rogers Method uses the following relations of gases: CH_4/H_2 ; C_2H_2/CH_4; C_2H_4/C_2H_6; C_2H_2/C_2H_4. In the Tables I and II are presented the relations between gases and the codification related to fault type of the Rogers Method.

TABLE I
RATIO KEY GASES – ROGERS METHOD

Gases Ratio	Range Gases Ratio	Code
CH_4/H_2	< 0.1	5
	0.1 to1.0	0
	1.0 to 3.0	1
	> 3.0	2
C_2H_2/CH_4	< 1.0	0
	> 1.0	1
C_2H_4/C_2H_6	< 1.0	0
	1.0 to 3.0	1
	> 3.0	2
C_2H_2/C_2H_4	< 0.5	0
	0.5 to 3.0	1
	> 3.0	2

In the table II, N = Normal State; DP = Discharge Partial; OH1 = Overheating Temperature < 150º C; OH2 = Overheating Temperature between 150 and 200º C; OH3 = Overheating Temperature between 200 and 300º C; OHC = Overheating Conductor; CCW = Winding Circulating Current; CCTC = Tank and Core Circulating Current; FL = Flashover; AR = Arcing; CT = Sparking; DPCO = Partial Discharge with Tracking (note CO).

TABLE II
FAULTS DESCRIPTION TABLE – ROGERS METHOD.

CH_4/H_2	C_2H_2/CH_4	C_2H_4/C_2H_6	C_2H_2/C_2H_4	Diagnostic
0	0	0	0	N
5	0	0	0	DP
1 or 2	0	0	0	OH1
1 or 2	1	0	0	OH2
0	1	0	0	OH3
0	1	0	0	OHC
1	0	1	0	CCE
1	0	2	0	CCTC
0	0	0	1	FL
0	0	1 or 2	1 or 2	ARC
0	0	2	2	CT
5	0	1 or 2	1 or 2	DPCO

Doernenburg Method is also based in ratio between gases, but, unlikely Roger's Method, its ratio between gases is not codified. In this method, the relative value of gases is associated with the fault conditions, in accordance with the Table III. In the Doernenburg Method is not considered the normal state, but there are restrictive conditions for application of this method, as showed in the Table IV. If at least one of the gases in each ratio between gases exceeds the limit L1, the ratio key gases procedure is valid, otherwise, if it is not significant d the unit should be resample and investigated by alternative procedures.

TABLE III
RATIO KEY GASES – DOERNENBURG METHOD.

Fault	Range Gases Ratio		Range Gases Ratio	
	CH_4/H_2	C_2H_2/C_2H_4	C_2H_6/C_2H_2	C_2H_2/CH_4
TD	> 1	< 0,75	> 0,4	< 0,3
DPL	< 0,1	NS	> 0,4	< 0,3
DPH	< 1 to > 0,1	> 0,75	< 0,4	> 0,3

where: NS = Not Significant; TD = Thermal Decomposition; DPL = Partial Discharge (Low-Intensity); DPH = Partial Discharge (High-Intensity).

TABLE IV
LIMIT CONCENTRATIONS OF DISSOLVED GAS [PPM].

Gas	H_2	CH_4	CO	C_2H_2	C_2H_4	C_2H_6
L1	100	120	350	35	50	65

The concentration in "parts per million" (ppm) of CH_4, C_2H_4 and C_2H_2 in the Duval Triangle Method are expressed as percentages of the total ($CH_4 + C_2H_4 + C_2H_2$) and plotted as a point (% CH_4, % C_2H_4, % C_2H_2) in a triangular coordinate system chart (Figure 1), which has been subdivided into fault zones. The fault zone in which the point is located designates the likely fault type produced by combination of gas concentration.

In the Duval Triangle (Figure 1), PD = Discharge Partial; T1 = Thermal Fault, Temperature < 300º C; T2 = Thermal Fault, 300 º C < Temperature < 700º C; T3 = Temperature >700º C; D1 = Discharges of Low Energy; D2 = Discharges of High Energy; DT = Mixture of Electrical and Thermal Faults.

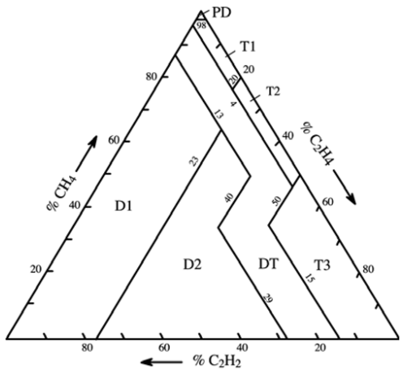


Figure 1. Duval Triangle Method.

The fault diagnostic methods of Rogers, Doernenburg and Duval Triangle as presented previously, are based in the ratio gases. These methods are the basis of the IEEE, IEC e ABNT Methods [9], as showed in the Tables V and VI.

TABLE V
RATIO KEY GASES - IEEE, IEC AND ABNT METHODS.

Range	Code		
	CH_4	C_2H_2/C_2H_4	C_2H_4/C_2H_6
< 0.1	0	1	0
0.1-1	1	0	0
1-3	1	2	1
> 3	2	2	2

TABLE VI
FAULTS DESCRIPTION TABLE - IEEE, IEC AND ABNT METHODS.

Diagnostic	Code		
	CH_4	C_2H_2/C_2H_4	C_2H_4/C_2H_6
N	0	0	0
DPL	0	1	0
DPH	1	1	0
D1	1	0	0
D2	1	0	0
T1	0	1	0
T4	0	1	0
T2	0	0	1
T3	1	0	1

where: DPL = Partial Discharge (Low Energy); DPH = Partial Discharge (High Energy); T4 = Thermal Fault, 150 º C < Temperature < 300º C.

III. INTELLIGENT SYSTEMS FOR FAULTS DIAGNOSING

The conventional methods of interpretation of the DGA are easy implemented and present good results for the hard faults prevention or after the occurrence of these faults, but they will have a little sensibility in relation to incipient faults. In spite of, DGA conventional methods can be used as standards for the intelligent systems based in the artificial intelligence [2, 8-12]. The intelligent systems applications for the incipient fault diagnostic in transformers are useful, because the intelligent systems are adaptative, capable of handling highly non-liner relationship and also generalize solutions for a new solution.

III. 1. ARTIFICIAL NEURAL NETWORK

ANN has been established as a useful tool for classification and regression problems, mainly for pattern recognitions and function approximations. An important characteristic of the ANN is that is not necessary to obtain a complete knowledge about the relations among the variables involved in the problem.

The main characteristics of the ANN are [6]:

- The basic information-processing unit of the ANN is called neuron;
- Due to complexity of the problems (classification or regression) the neurons are organized in three or more layers: the input layer, the output layer and one or several hidden layers;
- Each neuron receives input signals, which are multiplied by synaptic weights;
- Input signals are transformed into signal output to the next neuron and so on, by an activation function;
- The synaptic weights in the network are adjusted according to external signals received by neural network training process.

III.2. LEAST SQUARES SUPPORT VECTOR MACHINES

Least Squares Support Vector Machines (LS-SVM) is a method used for solving non-linear classification or modeling problems and has been applied to classification, function estimation and nonlinear system optimal control problems. The basis of the method is the mapping of all available data points to a feature space, thus transforming the problem into a simple linear problem. LS-SVM is an extension of the standard SVM and it expresses the training in terms of solving a linear set of equations instead of a quadratic programming problem, which allows LS-SVM to be faster than SVM and provides the reduced computation time. Efficient and scalable algorithms, such as those based on conjugate gradient can be applied to solve LS-SVM. Extensive studies have shown that LS-SVM is comparable to SVM in terms of generalization performance.

IV. METHODOLOGY

In these paper is proposed a tool for power transformers classification in relation to their failure condition, using dissolved gas analysis.

DGA methods, ANN and LS-SVM approaches were studied and implemented for verification of operation condition of power transformer. The data of this study were obtained from electricity utility, and each method was tested with all the data set.

IV. 1. ANN IMPLEMENTATION

ANN used is a multi-layer perceptron (MLP) with 5 neurons in the input layer, 12 neurons in the intermediate layer and 3 neurons in the output layer. The activation functions used in each of the layers were, respectively, tangent sigmoid function, tangent sigmoid function and logarithmic sigmoid function. MLP ANN doesn't presents binary output, so it was used the winner neuron strategy to make the classification. As the MLP used is composed of 3 output neurons, then it was determined that: 1. the output pattern of the neuron that to present higher value is defined as 1; 2. the output pattern of the remaining neurons is defined as 0. Because designing an innovative ANN learning procedure is beyond the scope of this paper, the ANN used in this paper was implemented using Neural Toolbox MATLAB routines [13].

IV. 2. LS-SVM IMPLEMENTATION

Classification with the LS-SVM approach was made with two networks. The first network was used to determine the transformer actual operation condition (normality or fault) and the second network was used to determine the fault type (heating or electric failure). Because designing an innovative LS-SVM learning procedure is beyond the scope of this paper, the LS-SVM network used in this paper was implemented using the LS-SVMlab Toolbox MATLAB routines [14].

V. PRESENTATION AND ANALYSIS OF RESULTS

In order to training of the ANN and LS-SVM were used in this study 246 samples of DGA and were selected and used data of 30 transformers for ANN and LS-SVM testing. The same data were used for the comparison between the results of the traditional methods, ANN and LS-SVM. In this study were used five key gases, all of them combustible, as inputs features of networks: H_2, CH_4, C_2H_2, C_2H_4 e C_2H_6. Overheating, electric failure or normal are the output features. Then, input and output patterns for each diagnosis criterion are defined according each method (Doernenburg Ratio Method, Rogers Ratio Method, IEC 60599, IEEE C57.104-1991 and Duval Triangle. The concentration of gases data in ppm are shown in the Table VII.

Table VIII shows the accuracy results for the incipient fault detection of the DGA conventional methods, ANN and LS-SVM. It is demonstrated in the Table VIII the superior efficiency of the ANN methods over the other methods. Rogers's method presented 33.33% efficiency, Dorneburg's method presented 50%, NBR7274 presented 50%, ANN presented 83.33% and LS-SVM presented 70%. These results demonstrate that intelligent systems (ANN and LS-SVM) can be a promising tool in the detection of incipient faults in power transformers.

The resulting tests employing the DGA conventional methods, ANN and LS-SVM approaches, are shown in the Table IX, where: TR = Transformer; D = Diagnostic; N = Normal State; OH = Overheating; FE = Electric Failure; LS = LS-SVM; NI = Fault is Unidentifiable; R = Rogers Method; DV = Duval Method; DB = Doernenburg; Method; NBR = NBR 7274 Method; IEC = IEC Standard 6059 Method.

TABLE VII
CONCENTRATION OF GASES IN PPM.

TR	Concentration of Gases [ppm]				
	H_2	CH_6	C_2H_4	C_2H_2	C_2H_6
1	40	62	11	1	51
2	600	1800	3800	130	520
3	6	2	7	0,4	5
4	1230	163	233	692	27
5	630	670	1100	1700	81
6	20	3	5	0,4	2
7	24	3	8	0,4	4
8	30	520	1000	0,4	310
9	645	86	110	317	13
10	110	860	1500	1800	300
11	19	9	7	0,4	20
12	16	1	1	0,4	1
13	1600	55000	74000	9300	42000
14	95	10	11	39	0,4
15	200	110	150	230	12
16	13	10	55	0,4	3
17	48	12	18	0,4	14
18	44	95	13	0,4	33
19	595	80	89	244	9
20	94	5	25	20	4
21	290	16	12	0,4	24
22	56	6	7	10	5
23	420	1400	1500	7	640
24	1790	580	336	919	321
25	0	1200	1900	4400	130
26	14	3	12	0,4	3
27	8800	64064	9565	0,4	72128
28	34	41	5	0,4	69
29	1330	10	66	182	20
30	230	120	220	740	14

TABLE VIII
DIAGNOSIS ACCURACY RESULTS.

Method	Diagnosis Accuracy [%]
Rogers	33.33
Duval	40.00
Dornenburg	50.00
NBR 7274	50.00
IEC 60599	50.00
LS-SVM	70.00
MLP	83.33

VI. CONCLUSION

A comparative study was made employing intelligent systems (ANN and LS-SVM) and DGA conventional methods (Doernenburg Ratio Method, Rogers Ratio Method, IEC 60599, IEEE C57.104-1991 and Duval Triangle) for the efficiency calculation for incipient fault detection on power transformers. The MLP and LS-SVM were trained with the gases relation of the DGA conventional methods cited previously.

TABLE IX
RESULTS SET DISPOSED BY DGA CONVENTIONAL METHODS, ANN AND LS-SVM.

TR	D	METHODS						
		R	DV	DB	NBR	IEC	MLP	LS
1	N	SA	NI	N	OH	OH	OH	OH
2	N	NI	NI	OH	OH	OH	OH	FE
3	OH	NI	FE	N	OH	OH	N	N
4	FE	FE	FE	NI	FE	FE	FE	FE
5	FE	NI	FE	NI	NI	NI	FE	FE
6	N	OH	FE	N	OH	OH	N	N
7	OH	NI	FE	N	OH	OH	N	N
8	FE	NI	FE	OH	OH	OH	OH	FE
9	FE	FE	FE	NI	FE	FE	FE	FE
10	N	NI	FE	NI	NI	NI	FE	FE
11	N	OH	FE	N	N	N	N	N
12	OH	NI	FE	N	NI	NI	N	N
13	FE	NI	FE	OH	NI	NI	OH	FE
14	FE	FE	FE	N	FE	FE	FE	FE
15	N	FE	FE	NI	FE	FE	FE	FE
16	N	NI	FE	N	NI	NI	N	N
17	OH	NI	FE	N	OH	OH	N	N
18	FE	OH	FE	N	OH	OH	N	OH
19	FE	FE	FE	NI	FE	FE	FE	FE
20	N	NI	FE	N	NI	NI	FE	FE
21	N	NI	FE	N	FE	FE	N	FE
22	OH	FE	FE	N	FE	FE	FE	FE
23	FE	NI	FE	OH	OH	OH	OH	FE
24	FE	FE	FE	NI	FE	FE	FE	FE
25	N	FE	FE	NI	FE	FE	FE	FE
26	OH	NI	FE	N	NI	NI	N	N
27	OH	NI	FE	OH	OH	OH	OH	FE
28	OH	OH	FE	N	OH	OH	OH	OH
29	FE	NI	FE	FE	NI	NI	FE	FE
30	FE	FE	FE	NI	FE	FE	FE	FE

The main problem of the DGA conventional methods is that different methods applied to the same sample result in different and often in contrary diagnostic decisions. An adequate utilization of the DGA conventional methods is important; because most of expertise's in incipient fault diagnostics use these methods for elaboration of expertness of power transformers operation condition. The diagnostic made using DGA conventional methods often are based in the empiric knowledge of the expertises, and then it is necessary to mind mainly with the inconsistency (non-decision problems) and the conflicts between the methods.

ANN due to its generalization characteristic is an adequate tool for to solve the conflicts between methods. The results presented show the actual capacity of the ANN for the identification of incipient fault in power transformers and also to solve the standardized inconsistencies of the methods. In spite of the superior results obtained with MPL, it was verified that with the default parameters of the respective algorithms the implementation of the LS-VM model is easier than the MLP model. MLP network involves more experience for its modeling and training, mainly for the definition of the number of hidden layers. Therefore, LS-SVM would be studied as an important alternative to ANN and DGA conventional methods for the identification of incipient fault in power transformers. We intend to continue the studies on the application of LS-SVM in modeling and simulation for the identification of incipient fault in power transformers, because it showed to be a useful tool in this study.

ACKNOWLEDGMENT

The authors acknowledge CNPQ – Brazil, for the financial support.

REFERENCES

[1] Assunção, T. C. B. N. Contribuição à Modelagem e Análise do Envelhecimento de Transformadores de Potência, *Tese de Doutorado*, Universidade Federal de Minas Gerais, UFMG, 2007.

[2] Singh, J., Sood, Y. R., Jarial, R. K., Verma, P., Condition Monitoring of Power Transformers - Bibliography Survey, *IEEE Electrical Insulation Magazine*, vol. 24, n° 3, May/June 2008, pag. 11-25.

[3] IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers, *C57.104-1991*, 1992.

[4] Mineral Oil-Impregnated Electrical Equipment in Service, Guide to the Interpretation of Dissolved and Free Gases Analysis, *IEC Standard*, 1999.

[5] Interpretação da Análise dos Gases de Transformadores em Serviço, *NBR 7274*, ABNT, 1982.

[6] Haykin, S. ; *Redes Neurais Princípios e prática*, 2ª edição, Prentice Hall INC., ARTMED Editora S. A., Bookman, 1999.

[7] Suykens, J. A. K., L., L., Doren, P. V. e Moorand, B., V. J., Least squares support vector machines classifiers: A large scale algorithm·, *European Conference on Circuits Theory and Design*, pp. 839-842.

[8] Moreira, M. P., Santos, L. T. B., Vellasco, M. M. B. R., Power Transformers Diagnosis Using Neural Networks, *Proceedings of International Joint Conference on Neural Networks*, Orlando, Florida, August 2007..

[9] Guardado, J. L., Naredo, J. L., Moreno, P., Fuerte, C. R., A Comparative Study of Neural Network Effciency in Power Transformers Diagnosis Using Dissolved Gas Analysis, *IEEE Transactions on Power Delivery*, vol. 16, nº 4, October 2001, pag. 643-647.

[10] Yanming, T., Zheng, Q., DGA Based Insulation Diagnosis of Power Transformers Via ANN, *Proceedings of The 6 th International Conference on Properties and Applications of Dielectric Materials*, June 2000, Xi'an Jiatong University, China.

[11] Farag, A. S., Mohandes, M., Al-Shaikh, A., Diagnosing Failed Distribution Transformers Using Neural Networks, *IEEE Transactions on Power Delivery*, vol. 16, n° 4, October 2001, pag. 631-636.

[12] Salami, A., Pahlevani, P., Neural Network Approach for Fault Diagnosis Transformers, *2008 International Conference on Condition Monitoring and Diagnosis, Beijing,* China, April 2008.

[13] Demuth, M. B. H. e Hagan, M., *Neural network toolbox, for use with MATLAB*, The Math Works;

[14] Pelckmans, K.; Suykens, J. A. K, *LS-SVMlab toolbox*, 2003, Version 1.5.

Modelling Control of Pore Number and Radii Distribution in Single-Cell Electroporation

Sadhana Talele Department of Engineering
University of Waikato
Hamilton, New Zealand
Email: sadhana@waikato.ac.nz
Paul Gaynor Department of Electrical and Computer Engineering
University of Canterbury
Christchurch, New Zealand

***Abstract*—Electroporation EP, in which external electric field pulses create transient pores in a cell membrane, is an important technique for delivery of genes and drugs into the cell. To enable a useful level of entry of genes into cells, the pores should have sufficiently large radii, and remain open long enough without causing membrane rupture. A numerical model for a single spherical cell electroporated by application of direct and/or alternating external electric field pulses has been developed. The model is used to indicate the actual number of pores and their radii distribution developed in response to various electric field pulses, as function of time and position on the cell surface. This study briefly describes the model briefly which is then used to investigate the ability to control the number and distribution of pore radii by choice of electric field parameters. We believe this would be one of the first papers to investigate the ability to CONTROL the number and distribution (range) of pore radii (as opposed to other papers that merely report the pore number and range with varying pulse parameters.**

***Index Terms*—electroporation, numerical model, fractional pore area, pore radii**

I. Introduction

Exposure of biological cells to electric fields can lead to a variety of biophysical and biochemical responses [1]–[3]. Electroporation (EP) is a feature common to all lipid bilayers in which electrically conductive pores are induced to form in the membrane, when the transmembrane potential V_m exceeds a semi-critical value creating a state of high permeability [4], [5]. These pores shunt excess stimulus current across the membrane, limiting the growth of V_m. The pores are long lived, sometimes surviving in the membrane for up to several minutes, and providing pathways for the movement of large ions and high molecular weight molecules such as drugs and DNA into the cell [6]–[8]. These properties have resulted the application called EP which is now a common tool in biotechnology and is used in a number of medical treatments [1]–[3], [6], [9]. However, the process of EP is still not completely understood. There is a need for supplementing experimental knowledge with theoretical models [10]. Depending on the type of application of EP it would be good to know the starting parameters of electric field to be applied in order to achieve a good number of pores in a particular range of size, as pores that are too small are not of any therapeutic use and reduce the strength of the cell membrane, while pores that are too big create a possibility of them becoming permanent leading to cell lysis. To date, such controllability of pore number and size has not been reported in the literature.

Previous reported studies have modelled EP to calculate V_m and the pore density N [10], [11]. These models consider the non-linear behavior of EP, which is due to the dynamics of membrane pore formation and its effect on the electrical properties of the membrane in turn. However, these models do not include the evolution of pore sizes. They assume a constant pore radius. Recent models [12], [13] include spatial and temporal aspects of pore radius evolution. However, the form of electric fields used in these models were single unipolar pulses. The study presented here develops a model of single spherical cell EP and provides simulation of spatial and temporal aspects of pore radius as an effect of multiple pulses of varied magnitudes. As such it is possible to investigate potential control of both the number of pores created and their radii distribution.

II. Model of a single cell

Consider a spherical cell of radius a with intracellular conductivity σ_{in}. The cell is immersed in a medium with conductivity σ_{ex}. This system is exposed to a time-varying electric field $E(t)$. Azimuthal symmetry about the axis of applied electric field is assumed. The model discussed in [14] is described below briefly for completeness.

A. Transmembrane potential

It is assumed that the intracellular and extracellular regions are charge free. The homogeneous external electric field of strength

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_40, © Springer Science+Business Media B.V. 2010

$E(t)$ is used as a boundary condition, the potential being fixed at

$$\Phi(3a,\theta) = -3aE(t)\cos\theta, \tag{1}$$

on a sphere of radius $3a$ surrounding the cell, where θ is the polar angle. Since there are no sources inside the region, the potential obeys Laplace's equation

$$\nabla^2\Phi = 0, \tag{2}$$

except at the cell membrane at a radial distance of $(r = a)$ where the potential is discontinuous because of the abrupt change in conductivity and the approximated infinitely thin membrane. Thus internal and external potentials can be defined as,

$$\Phi(r,\theta) = \begin{cases} \Phi_{\text{in}}(r,\theta) & r < a, \\ \Phi_{\text{ex}}(r,\theta) & a < r < 3a. \end{cases} \tag{3}$$

The current across the membrane is used to relate the internal and external potentials [10], that is,

$$-\hat{\mathbf{r}} \cdot (\sigma_{\text{in}}\nabla\Phi_{\text{in}}(a,\theta)) = -\hat{\mathbf{r}} \cdot (\sigma_{\text{ex}}\nabla\Phi_{\text{ex}}(a,\theta)) = C_m\frac{\partial V_m}{\partial t} + J_m \tag{4}$$

where

$$V_m(\theta) = \Phi_{\text{in}}(a,\theta) - \Phi_{\text{ex}}(a,\theta). \tag{5}$$

Here $\hat{\mathbf{r}}$ is the unit outward radial vector, C_m is the specific membrane capacitance, V_m is the transmembrane potential and J_m is the current density at the cell membrane due to existing pores. The current density is made up of three terms as given in Equation 6

$$J_m = J_{\text{ion}} + J_{\text{sml}} + J_{\text{lge}}, \tag{6}$$

where $J_{\text{ion}} = g_m(V_m - V_{\text{rest}})$ is the ionic current density [10] (g_m is the specific membrane conductance, and V_{rest} is the membrane rest potential). The remaining two terms are explained in the following paragraphs.

J_{sml} is the current density through small pores and is given by

$$J_{\text{sml}} = Ni_{\text{sml}}(r), \tag{7}$$

where N is the pore density of the initial small pores formed and $i_{\text{sml}}(r)$ is the diffusion current through a single pore of radius r (true for small pores only). A previously derived expression for i_{sml} [10], based upon the Nernst-Planck equation models, is used for pore radius below 1 nm, and is,

$$i_{\text{sml}} = \frac{\pi r^2\sigma_{\text{ps}}\nu_m RT}{Fh} \cdot \frac{(e^{\nu_m} - 1)}{(G_- e^{\nu_m} - G_+)} \tag{8}$$

with

$$G_\pm = \frac{w_0 e^{w_0 \pm n\nu_m} \pm n\nu_m}{w_0 \pm n\nu_m}. \tag{9}$$

Here σ_{ps} is the conductivity of the aqueous solution that fills the pore (approximated by $\sqrt{\sigma_{\text{in}}\ \sigma_{\text{ex}}}$), F is Faraday's constant, R is the universal gas constant, T is the absolute temperature, h is the thickness of the membrane, w_0 is the energy barrier inside the pore, n is the relative entrance length of the pore and v_m is the nondimensional transmembrane potential [10] given by

$$v_m = V_m\left(\frac{F}{RT}\right). \tag{10}$$

This equation for pore current i_{sml} (equation 8) accounts for the electrical interactions between the ions and the pore wall [10]. Now, assume Q larger pores exist, and i_{lge} is the current through the electropores of radius larger than 1 nm. Then, J_{lge} is the total current density through Q larger pores; r_q being the radius of the q^{th} pore. Hence,

$$J_{\text{lge}} = \frac{1}{A}\sum_{q=1}^{Q} i_{\text{lge}}(r_q) \tag{11}$$

where A is the corresponding cell surface area. For these larger pores, the current–voltage relationship assumes that the transmembrane potential V_m occurs across the sum of pore resistance R_p and the series input resistance R_{in} [13], [15], as follows:

$$i_{\text{lge}}(r) = \frac{V_m}{R_p + R_{\text{in}}}, \tag{12}$$

where

$$R_p = \frac{h}{\pi\sigma_{\text{ps}}r_q^2}, \tag{13}$$

and

$$R_{\text{in}} = \frac{1}{2\sigma_{\text{ps}}r_q}. \tag{14}$$

B. Formation of pores:

Initially pores are assumed to be formed with the minimum-energy radius r_m = 0.76 nm, at a rate given by [10]

$$\frac{dN}{dt} = \psi e^{(V_m/V_{\text{ep}})^2}\left(1 - \frac{N}{N_{\text{eq}}(V_m)}\right), \tag{15}$$

where N is the pore density of the initial small pores formed, ψ is the creation rate coefficient and N_{eq} is the equilibrium pore density for a voltage V_m given by

$$N_{\text{eq}}(V_m) = N_0 e^{b(V_m/V_{\text{ep}})^2}. \tag{16}$$

Here, N_0 is the initial pore density with no applied electric field, V_{ep} is the characteristic voltage of electroporation and b is the pore creation constant equal to $(r_m/r_*)^2$, where r_* is the minimum radius of hydrophilic pores and r_m the minimum-energy radius [16]. All parameter values are as given in Table 1.

C. Evolution of pore radii:

The pores that are initially created with minimum-energy radius r_m change in size. Evolution of the radius of a pore is governed by the following equations 17 - 21 [13]. The lipid bilayer energy w_m depends on a number of parameters and is given by

$$w_m = \sum_{q=1}^{Q} \left[w_{\rm st} \left(\frac{r_*}{r_q} \right)^4 + 2\pi w_{\rm ed} r_q - \pi \xi_{\rm eff}(A_p) r_q^2 + \int_0^{r_q} F(r_q, V_m) dr \right]. \quad (17)$$

The terms in this equation are explained in the following paragraphs.

In the first term, $w_{\rm st}$ is the steric repulsion energy [16]. The lipid head groups which line the pore interior, tend to repel each other due to steric and/or electrostatic interactions [17]–[19] and are taken into account by this term in equation 17.

The appearance of a circular pore in a membrane is balanced by the presence of two energy terms: reduction in energy barrier proportional to removal of pore area $\pi {r_q}^2$ (third term in equation 17) and increase in energy barrier by a linear edge component proportional to pore edge of length $2\pi r_q$ [4], [20], [21] (second term in equation 17). Here, $w_{\rm ed}$ is the pore edge energy, and $\xi_{\rm eff}$ the effective tension of the membrane is given by

$$\xi_{\rm eff}\,(A_p) = 2\xi' - \frac{2\xi' - \xi_0}{\left(1 - \frac{A_p}{A}\right)^2} \quad (18)$$

where ξ' is the energy per area of the hydrocarbon-water interface [16], [18], ξ_0 is the tension of a membrane without pores and A is the total area of the lipid bilayer. A varying value of the total area A_p occupied by the pores at any given time is given by

$$A_p = \sum_{q=1}^{Q} \pi r_q^2 \quad (19)$$

and contributes to a dynamic effect of pore formation both temporally and spatially.

The last term in Equation 17 is the contribution of the membrane potential to the bilayer energy [16]. Assuming the inner surface of a pore as toroidal [16], [22], the electric force F acting on the pore is given by

$$F(r, V_m) = \frac{F_{max}}{\left(1 + \frac{r_h}{r + r_t}\right)} V_m^2. \quad (20)$$

This equation is a heuristic approximation [16] of the numerical solution which has been computed for the electrical force acting on a pore derived from first principles; Here, r_h and r_t are constants taken from reference [16]. This equation is thought to be appropriate for larger pores as it predicts that F approaches a constant value $F_{\rm max}$ as the pore radius increases, rather than increase linearly [21], or decrease to zero [23]–[25] as radius increases. All parameter values are as given in Table 1.

a	15.0 (μm)	cell radius
C_m	10^{-2} (Fm^{-2})	specific membrane capacitance [12]
h	5.0 (nm)	membrane thickness [?], [5], [10]
g_m	1.9 (Sm^{-2})	specific membrane conductance [10]
$V_{\rm rest}$	–80 (mV)	membrane rest potential [10]
$\sigma_{\rm in}$	0.3 (S m^{-1})	intracellular conductivity [5]
$\sigma_{\rm ex}$	1.2 (Sm^{-1})	extracellular conductivity [?]
r_*	0.51 (nm)	minimum radius of hydrophilic pores [13]
r_m	0.8 (nm)	minimum energy radius at V_m=0 [13]
T	295 (K)	absolute room temperature [10]
n	0.15	relative entrance length of pores [10]
b	2.46	pore creation constant [10]
$V_{\rm ep}$	258 (mV)	characteristic voltage of electroporation [10]
N_0	1.5×10^9 (m^{-2})	initial pore density [10]
w_0	2.65	energy barrier within pore [10]
ψ	1×10^9 (m^{-2} s^{-1})	creation rate coefficient [10]
$w_{\rm st}$	1.4×10^{-19} (J)	steric repulsion energy [13]
$w_{\rm ed}$	1.8×10^{-11} (J m^{-1})	edge energy [13]
ξ_0	1×10^{-6} (J m^{-2})	tension of the bilayer without pores [12]
ξ'	2×10^{-2} (J m^{-2})	tension of hydrocarbon-water interface [10]
$F_{\rm max}$	0.70×10^{-9} (N V^{-2})	max electric force for V_m = 1 V [10]
r_h	0.97×10^{-9} (m)	constant [13]
r_t	0.31×10^{-9} (m)	constant [13]
D	5×10^{-14} (m^2 s^{-1})	diffusion coefficient for pore radius [12]

TABLE I
GEOMETRIC, ELECTRIC AND ELECTROPORATION PARAMETERS USED IN SIMULATION.

The rate of change of pore radii is given by [13]

$$\frac{dr_q}{dt} = -\frac{D}{kT}\frac{\partial w_m}{\partial r_q}, \qquad q = 1, 2, ..., Q \quad (21)$$

where D is the diffusion coefficient of the pore radius, k is Boltzman's constant and T is the absolute temperature.

III. NUMERICAL IMPLEMENTATION

The model described above is implemented in MATLAB and simulations carried out for a spherical cell with 15 μm radius. The following description of model geometry is illustrated in Figure 1. Azimuthal symmetry about the axis of the applied electric field is assumed. The membrane is divided into a number of individually modelled slices (25 including the two poles). Each section of cell membrane is assigned a pore density N, of 'small pores', idealised as having equal radius r_m. Each section also has a transmembrane voltage (V_m) calculated. All results are simulated for a radial discretization step of $3a/51$.

IV. SIMULATION AND RESULTS

The number of pores and their radius is significant when administering an application. It is important to select appropriate values of experimental electric field parameters in order that expected results occur. For example, a sufficient number of larger pores may be required if a particular gene transfer is expected. The model discussed above can simulate results of EP, given any form of applied electric field and most other EP system parameters. Of importance in this study is the effect of applied

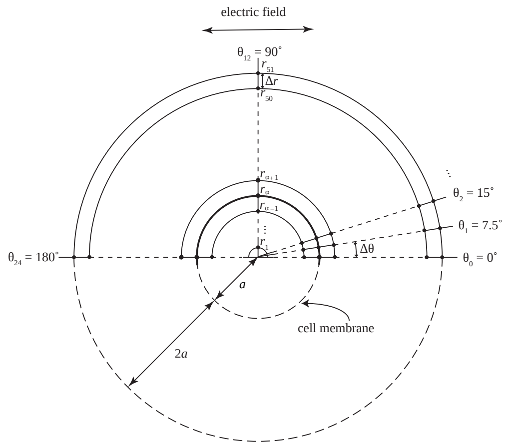


Fig. 1. Spherical cell model geometry showing the angle and radius discretization used in the numerical model

electric field on pore size distribution and the fraction of cell surface area occupied by pores, defined as the fractional pore area (*FPA*). Several simulations were carried out and important results tabulated as shown in Tables 2-5. These tables give information about the final pore numbers, pore density, %*FPA*, maximum mean radius, maximum overall radius, and the pore size distribution. The numbers on right hand side of the pore numbers is the %*FPA* contribution due to corresponding pores in that row. Electric field applied is represented as [135 0 135] for a time duration [0 3 6 9] . This represents an electric field pulse of 135 kV/m applied from 0 to 3 μs, a break from 3 to 6 μs, followed by an electric field of 135kV/m from 6 to 9 μs and so on. The magnitude of first pulse was chosen such that it is just enough to create pores resulting in an *FPA* of more than 0.03%, as the appropriate level of electroporation is in the range of 0.01% to 0.1% [26]. The magnitude and the time duration of second pulse were varied to investigate its effect on the number of pores in the range 10-20 nm. The model was used to investigate the mechanism /usefulness of the two-pulse protocol for control of pore size and number distribution. Different electric field amplitudes for different time durations were used as the second pulse for simulations, keeping other parameters like cell radius, internal and external conductivity, etc., constant. The final extent of electroporation (measured by *FPA*) was targeted to be about 0.05 % , with at least over one tenth of the *FPA* contributed by pores with radii larger than 5 nm, making sure that the number of these pores larger than 5 nm in radius is also at least 2% of the total number of pores on a spherical cell of radius 15 μm.

As seen in Table 2, the first pulse was chosen to be of amplitude 135 kV/m with a time duration of 3 μs, as this gave a reasonable *FPA* and a maximum pore radius of 10 nm. The aim is to increase the total *FPA*, as well as the contribution of *FPA* by pores in the range greater than 5 nm, preferably larger than 10 nm. Several amplitudes for the second pulse were tried (135kV, 175kV, and 195kV) with a break of 3 μs between two pulses.

As seen in Table 3, the second pulse of 195 kV/m, increases the total *FPA* substantially to 0.05%. The total number of pores remain nearly the same (a few new pores form as an effect of the second pulse), while increasing the number of pores in the range 5-10 nm from 660 to 862, pores in the range 10-20nm from 275 to 376 and pores in the range 20-100 nm from 0 to 14. As seen in Table 3, the *FPA* contribution due to large pores in the pore radius range 5-100 nm has increased from about 18% of total to about 27 % of total *FPA*.

In an attempt to increase this ratio, a higher amplitude second pulse was tried (results not shown). This, instead of increasing the size of pores, formed new smaller pores. Using a higher magnitude first pulse hasve been seen to give a higher total *FPA* but with more small pores (under 5 nm).

Another way of reaching the target is to use the pulses for a longer time. The time duration of the first pulse (135kV/m) and the second pulse (195kV/m) was increased to 4 μs, with different break times of 3 μs, 8 μs and 26 μs. The results are shown in Table 4 to see the effects on *FPA* and pore sizes. The total *FPA* increases in both cases, so does the contribution of *FPA* by pores larger than 5 nm. A shorter break ensures that the initial pores do not shrink too much and thus a few pores are able to expand to a size larger than 20 nm. It is not possible for these pores to exist with a longer break. However, they could be possible with longer first and/or second pulse duration.

To determine if use of a much lower amplitude first pulse or reversing the sequence of the higher and the lower amplitude pulses would have any favorable result, more simulations were done and results are shown in Table 5. As seen in the first column of Table 5, the number of large pores and their *FPA* contribution (radius larger than 5 nm) drastically reduces although the total *FPA* is similar. Column 4 of Table 5 indicates that while a very small (85 kV/m) first pulse followed by a high amplitude second pulse (195 kV/m) gives a reasonable value of total *FPA*, and a good number of large pores (5-10 nm), it does not support formation of many pores in the range (greater than 10 nm).

V. Conclusion

There seems to be an interdependent relationship between the pulse amplitude, pulse duration, number of pulses and the time-gap between pulses in order to control the total *FPA* and the desired pore radii distribution. Generally, the first pulse seems to control the number of pores, and the second pulse helps to grow these pores to a larger size. This is in agreement with the

Table 2

Time in microseconds	2us	2us	3us
Electric field Amplitude in Volts/m	125K	135k	135k
(total)Final pore number	**56246**	**78720**	**78720**
(total)Final FPA (%)	**0.02767**	**0.03345**	**0.03405**
(total)Maximum mean radius (nm)	2.0101	1.8779	1.8779
(total)Maximum overall radius (nm)	**8.3189**	**7.456**	**10.9606**

Table 3

Time in microseconds	**[0 3 6 9]us**		**[0 3 6 9]us**		**[0 3 6 9]us**	
Electric field Amplitude in Volts/m	**[135k 0 135k]**		**[135k 0 175k]**		**[135k 0 195k]**	
(total)Final pore number	**78720**		**78739**		**79011**	
(total)Final FPA (%)	**0.03539**		**0.04599**		**0.05257**	
(total)Maximum mean radius (nm)	1.8779		2.0671		2.1638	
(total)Maximum overall radius (nm)	**18.6556**		**25.1665**		**25.9748**	
pores less than 1 nm	1465	0.00009	1468	0.00009	1639	0.00011
pores in the range 1-5 nm	76320	0.029	76121	0.03586	76120	0.03873
pores in the range 5-10 nm	660	0.0023	861	0.00421	862	0.00487
pores in the range 10-20 nm	**275**	**0.0035**	**282**	**0.00544**	**376**	**0.0081**
pores in the range 20-100 nm	0	0	7	0.00039	14	0.00076
pores larger than 100 nm	0	0		0	0	0

Table 4

Time in microseconds	**[0 4 7 11]**		**[0 4 12 16]**		**[0 4 30 34]**	
Electric field Amplitude in Volts/m	**[135k 0 195k]**		**[135k 0 195k]**		**[135k 0 195k]**	
(total)Final pore number	**78850**		**83903**		**84191**	
(total)Final FPA (%)	**0.05573**		**0.05394**		**0.05444**	
(total)Maximum mean radius (nm)	2.1705		2.1838		2.1903	
(total)Maximum overall radius (nm)	**33.6247**		**22.7617**		**16.6944**	
pores less than 1 nm	1494	0.0001	1926	0.00012	1816	0.00012
pores in the range 1-5 nm	76120	0.03826	80120	0.0424	80320	0.04193
pores in the range 5-10 nm	861	0.00557	1779	0.00889	1781	0.00799
pores in the range 10-20 nm	**359**	**0.01035**	**65**	**0.00222**	**274**	**0.00441**
pores in the range 20-100 nm	16	0.00145	13	0.00066	0	0
pores larger than 100 nm	0	0	0	0	0	0

Table 5

Time in microseconds	**[0 4 7 11]**		**[0 3 6 9]**	
Electric field Amplitude in Volts/m	**[195000 0 135000]**		**[85000 0 195000]**	
(total)Final pore number	**211812**		**(pores form at second pulse)126748**	
(total)Final FPA (%)	**0.05313**		**0.05778**	
(total)Maximum mean radius (nm)	1.6206		2.7365	
(total)Maximum overall radius (nm)	**24.829**		**11.6053**	
pores less than 1 nm	6500	0.00062	1580	0.0001
pores in the range 1-5 nm	20500	0.04774	12310	0.0495
pores in the range 5-10 nm	40	0.00013	2060	0.00807
pores in the range 10-20 nm	**270**	**0.00442**	**10**	**0.0001**
pores in the range 20-100 nm	1	0.00023	0	0
pores larger than 100 nm		0	0	0

theoretical research by citeneu2003 and experimental findings by [27], who reported better uptake of DNA for a two pulse protocol compared with a single pulse protocol. The break between the two pulses will let the initial pores reduce in size. Thus if the break is too long, the second pulse tends to form new pores and has to expand the initial pores from a smaller size, thus ending up with a comparatively smaller size of pores than if a shorter break was used.

REFERENCES

[1] M. Cemaẑar, D. Miklavĉiĉ, and G. Serŝa, "Intrinsic sensitivity of tumor cells to bleomycin as an indicator of tumor response to electrochemotherapy," *Japanese Journal of Cancer Research*, vol. 89, pp. 328–333, 1998.

[2] G. Serŝa, "Electrochemotherapy," in *Electrochemotherapy, Electrogenetherapy, and Transdermal Drug Delivery : Electrically Mediated Delivery of Molecules to Cells (Methods in Molecular medicine)*, M. J. Jaroszeski, R. Heller, and R. Gilbert, Eds. Totowa, New Jersey: Humana Press, 2000, ch. 6, pp. 119–133.

[3] G. Serŝa, S. Novaković, and D. Miklavĉiĉ, "Potentiation of bleomycin antitumor effectiveness by electrotherapy," *Cancer Letters*, vol. 69, pp. 81–84, 1993.

[4] R. W. Glaser, S. L. Leikin, L. V. Chernomordik, V. F. Pastushenko, , and A. I. Sokirko, "Reversible electrical breakdown of lipid bilayers: Formation and evolution of pores," *Biochim Biophys Acta*, vol. 940, p. 275287, 1988.

[5] T. Kotnik, D. Miklavĉiĉ, and T. Slivnik, "Time course of transmembrane voltage induced by time-varying electric fields: a method for theoretical analysis and its application," *Bioelectrochemistry and Bioenergetics*, vol. 45, pp. 3–16, 1998.

[6] L. M. Mir, "Therapeutic perspectives of *in vivo* cell electropermeabilization," *Bioelectrochemistry*, vol. 53, pp. 1–10, 2000.

[7] T. Budak-Alpdogan, D. Banerjee, and J. R. Bertino, "Hematopoietic stem cell gene therapy with drug resistance genes: An update," *Cancer Gene Therapy*, vol. 12, pp. 849–863, 2005.

[8] J. R. Bertino, "Transfer of drug resistance genes into hematopoietic stem cells for marrow protection," *The Oncologist*, vol. 13, pp. 1036–1042, 2008.

[9] S. B. Dev, D. P. Widera, and G. A. Hofmann, "Medical applications of electroporation," *IEEE Transactions on Plasma Science*, vol. 28, pp. 206–223, 2000.

[10] K. A. DeBruin and W. Krassowska, "Modeling electroporation in a single cell. I. Effects of field strength and rest potential." *Biophysical Journal*, vol. 77, pp. 1213–1224, 1999.

[11] T. R. Gowrishankar and J. C. Weaver, "An approach to electrical modeling of single and multiple cells," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 3203–3208, 2003.

[12] W. Krassowska and P. D. Filev, "Modeling electroporation in a single cell," *Biophysical Journal*, vol. 92, pp. 404–417, 2007.

[13] K. C. Smith, J. C. Neu, and W. Krassowska, "Model of creation and evolution of stable electropores for DNA delivery," *Biophysical Journal*, vol. 86, pp. 2813–2826, 2004.

[14] Talele, S., P. Gaynor, J. van Ekeran, and M. J. Cree. Modelling single cell electroporation with bipolar pulse: Simulating dependence of electroporated fractional pore area on the bipolar field frequency. In: *Proceedings International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering*. USA (2008).

[15] J. Newman, "Resistance for flow of current to a disk," *Journal of The Electrochemical Society*, vol. 113, pp. 501–502, 1966.

[16] J. C. Neu and W. Krassowska, "Modeling postshock evolution of large electropores," *Physical Review E*, vol. 67, pp. 1–12, 2003.

[17] J. C. Weaver and Y. Chizmadzhev, "Theory of electroporation: a review," *Bioelectrochemistry and Bioenergetics*, vol. 41, pp. 135–160, 1996.

[18] J. Israelachvili, *Intermolecular and Surface Forces*. London, UK: Academic Press, 1992.

[19] J. C. Neu and W. Krassowska, "Asymptotic model of electroporation," *Physical Review E*, vol. 59, pp. 3471–3482, 1999.

[20] J. C. Weaver and K. T. Powell, "Theory of electroporation," in *Electroporation and Electrofusion in Cell Biology*, E. Neumann, A. E. Sowers, and C. A. Jordan, Eds. New York: Plenum Press, 1989, ch. 7, pp. 111–126.

[21] I. G. Abidor, V. B. Arakelyan, L. V. Chernomordik, Y. A. Chizmadzhev, V. F. Pastushenko, and M. R. Tarasevich, "Electric breakdown of bilayer membranes: I. The main experimental facts and their qualitative discussion," *Bioelectrochemistry and Bioenergetics*, vol. 6, pp. 37–52, 1979.

[22] M. Kanduŝer, M. Foŝnaria, M. Ŝentjurc, V. Kralj-Igliĉ, H. Hägerstrand, A. Igliĉ, and D. Miklavĉiĉ, "Effect of surfactant polyoxyethylene glycol $C_{12}E_8$ on electroporation of cell line DC3F," *Colloids and Surfaces A*, vol. 214, pp. 205–217, 2003.

[23] A. Barnett and J. C. Weaver, "Electroporation: A unified, quantitative theory of reversible electrical breakdown and mechanical rupture in artificial planar bilayer membranes," *Bioelectrochemistry and Bioenergetics*, vol. 25, pp. 163–182, 1991.

[24] R. P. Joshi, Q. Hu, K. H. Schoenbach, and H. P. Hjalmarson, "Improved energy model for membrane electroporation in biological cells subjected to electrical pulses," *Physical Review E*, vol. 65, no. 4, pp. 041 920–041 928, Apr 2002.

[25] V. F. Pastushenko and Y. A. Chizmadzhev, "Stabilization of conducting pores in BLM by electric current," *General Physiology and Biophysics*, vol. 1, pp. 43–52, 1982.

[26] M. Hibino, M. Shigemori, H. Itoh, K. Nagayama, and K. Kinosita, Jr., "Membrane conductance of an electroporated cell analyzed by submicrosecond imaging of transmembrane potential," *Biophysical Journal*, vol. 59, pp. 209–220, 1991.

[27] S. Ŝatkauskas, M. F. Bureau, M. Puc, A. Mahfoudi, D. Scherman, D. Miklavĉiĉ, and L. M. Mir, "Mechanisms of *in vivo* DNA electrotransfer: Respective contributions of cell electropermeabilization and DNA electrophoresis," *Molecular Therapy*, vol. 5, pp. 133–140, 2002.

Single Allocation P-Hub Median Problem to Monitor Land Borders by Using Unmanned Aircraft

Yasemin Akkuş, İnci Sarıçiçek
Eskişehir Osmangazi University,
Department of Industrial Engineering

Abstract - **During the last two decades, hub location problems have become popular with successful applications in transportation, telecommunication, networks, retailing and other logistic systems. In this paper we consider Turkey's land borders security. Because security is the one of the most important deal nowadays, countries are spending so much to prevent threats that may come from neighbor countries' borders. Also several geographical restrictions at our land borders cause insufficiency to monitor and to take the requiring data. We focus on the problem of the single allocation p-hub median that is addressed to monitor the movement at Turkey's land borders by using unmanned aircraft. In addition, aiming assignment of demand points that are defined as cities at land borders to hubs mathematical model is built and the results are evaluated. Model is encoded by using GAMS software and solved by CPLEX solver.**

Key Words- P-Hub Median Problem, Unmanned Aircraft

I. INTRODUCTION

Hubs form critical elements in many airline, transportation, postal and telecommunications networks. They are centralized facilities in these networks whose functions are to consolidate, switch and sort flows. Typical applications of hub location include airline passenger travel, telecommunication systems and postal networks [1]. The hub location problem is concerned with locating hub facilities and allocating demand nodes to hubs in order to route the traffic between origin–destination pairs.

Studies on the hub location problem often assume three things [2]:

1. The hub network is complete with a link between every hub pair;
2. There are economies of scale incorporated by a discount factor for using the inter-hub connections;
3. No direct service (between two non-hub nodes) is allowed

There are several kinds of hub location problems based on characteristics of a particular hub network. If the number of hub nodes is fixed to p, we are dealing with p-hub location problems. Each non-hub node can be allocated either to one (single allocation) or to more hubs (multiple allocations). Capacitated versions of hub problems may involve different kinds of capacity constraints [3].

In general, problems are handled in 4 types: ***p-Hub Median Problem; p-Hub Center Problem; Hub Covering Problem; Hub Location Problem with Fixed Costs.*** The objective of the p-hub median problem is to minimize the total transportation cost (time, distance, etc.) needed to serve the given set of flows, given n demand nodes, flow between origin–destination pairs and the number of hubs to locate (p) [4]. p-hub center problem is a minimax type problem. The p-hub center problem is to locate p hubs and to allocate non-hub nodes to hub nodes such that the maximum travel time (or distance) between any origin–destination pair is minimized [5]. Hub covering problem's aim is to maximize the covered area by the hubs obeying the maximum time bound on travel time [4]. In the p-hub location problem the fixed cost of opening facilities is disregarded. On the other hand, the simple plant location problem includes fixed facility costs. In 1992, O'Kelly introduces the fixed facility costs into a hub location problem and thereby making the number of hubs a decision variable [6].

In this study, we addressed to monitor the movement at Turkey's land borders by using unmanned aircraft. Within the scope of the problem, we consider single allocation p-hub median problem.

The research on the p-hub median problem with single assignment started with the works of O'Kelly (1987). O'Kelly (1987) presented the first mathematical formulation for the single allocation p-hub median problem as a quadratic integer program which minimizes the total network cost. This quadratic integer program is considered as the basic model for hub location problem. Campbell (1994) produced the linear integer programming formulation for the single allocation p-hub median problem. His formulation has (n^4+n^2+n) variables of which (n^2+n) are binary and it has (n^4+2n^2+n+1) linear constraints [2].

Ernst and Krishnamoorthy (1996) proposed a different linear integer programming formulation which requires fewer variables and constraints in an attempt to solve larger

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_41,

problems. They treated the inter-hub transfers as a multi commodity flow problem where each commodity represents the traffic flow originating from a particular node [2].

Sohn and Park (1997) studied the single allocation two-hub median problems. They transform the quadratic 0-1 integer program for single allocation problem in the fixed two hub system into a linear program and solve in polynomial time when the hub locations are fixed [7].

Ebery (2001) presented a formulation for the single allocation p-hub median problem with two or three hubs [8].

Yaman (2008) considered a version of this problem where service quality considerations are incorporated through delivery time restrictions. She proposed mixed integer programming models for these two problems and report the outcomes of a computational study using the CAB data and the Turkey data [9].

Various heuristic algorithms have been developed by O'Kelly (1987), Aykin (1990), Klincewicz (1991, 1992), Skorin-Kapov and Skorin-Kapov (1994), Smith et al. (1996), Campbell (1996), Ernst and Krishnamoorthy (1996), O'Kelly et al. (1996), Skorin-Kapov et al. (1996), Ernst and Krishnamoorthy (1998), Pirkul and Schilling (1998), Abdinnour- Helm (2001), Elhedhli and Hu (2005) and Kratica (2007) .

In this study, we addressed to monitor the movement at Turkey's land borders by using unmanned aircraft. Within the scope of the problem, weight factors of the possible hubs are determined; a mathematical model was built and solved to decide hub locations.

The remainder of this paper is organized as follows. In the next section, we introduce problem. The problem is solved by in Section 3. Finally, our results are summarized in Section 4.

II. *p*-HUB MEDIAN PROBLEM

Hubs are centralized facilities in these networks whose functions are to consolidate, switch and sort flows. Flow concentration and consolidation on the arcs that connect hub nodes (hub arcs) allow us to exploit transportation flow economies. It is also possible to eliminate many expensive direct connection arcs between origin destination pairs [1].

There are two basic types of hub networks: single allocation and multiple allocations (Fig. 1) [10]. They differ in how non-hub nodes are allocated to hubs. In single allocation, all the incoming and outgoing traffic of every demand center is routed through a single hub; in multiple allocation, each demand center can receive and send flow through more than one hub [2].

The p-hub median problem is to locate p hubs in a network and allocate demand points to hubs such that the sum of the costs of transporting flow between all origin destination pairs in the network is minimized [4]. So p-hub median problem is a minimum type problem.

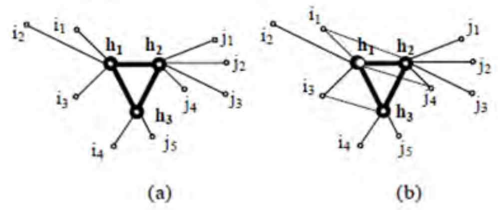


Fig. 1. (a)Single allocation (b) and multiple allocation

p-hub median problems can be classified into two groups according to the connection type of demand points to hubs as single and multi allocation. If each demand point is assigned to exactly one hub, the problem is called the single assignment (allocation) problem. If a demand point can be assigned to more than one hub then the problem is called the multi assignment problem. Our problem is the single assignment (allocation) p-hub median problem.

III. SINGLE ALLOCATION *P*-HUB MEDIAN PROBLEM

A. The Definition of The Problem

The land borders of the republic of Turkey are 2,573 km in total, and coastlines (including islands) are another 8,333 km, Turkey has two European and six Asian countries for neighbors along its land borders.

The land border to the northeast with the commonwealth of Independent States is 610 km long; that with Iran, 454 km long, and that with Iraq 331 kilometers long. In the south is the 877 km-long border with Syria, which took its present form in 1939, when the Republic of Hatay joined Turkey. Turkey's borders on the European continent consist of a 212-km frontier with Greece and a 269-km border with Bulgaria.

The border of Turkey with Syria, Iran, and Iraq are especially important when Turkey's geo-strategic location, its difficult neighborhood, the closeness to the energy sources in the Middle East, international transportation, and terrorist incidents in the east and southeast are considered. There is no physical security system at these borders except with Syria [11].

Nowadays, developed countries use state-of-the-art technologies to protect their borders such as Unmanned Aerial Vehicle (UAV), satellite-based navigation system, and sensors. With this study, the handiness of the UAV will be investigated for Turkey's borders.

Unmanned Aerial Vehicles (UAVs) are remotely piloted or self-piloted aircraft that can carry cameras, sensors, communications equipment or other payloads. They have been used in a reconnaissance and intelligence-gathering role since the 1950s, and more challenging roles are envisioned, including combat missions [12].

UAVs will make significant contributions to the war fighting capability of operational forces. They greatly improve the timeliness of battlefield information while reducing the risk of capture or loss of manned RECCE assets.

When compared to manned RECCE aircraft they are cost effective and versatile systems. While reconnaissance, intelligence, surveillance, and target acquisition (RISTA) are the premier missions of UAVs, they can also provide substantial support to intelligence preparation of the battlefield (IPB), situation development, battle management (BM), battle damage assessment (BDA) [13].

Each Region has its own unique regional features. The features of all the possible hubs and demand nodes are not assumed the same in order to have a realistic model. So, parameters that present the regional features are defined and used to reflect the differences between the regions. We used these features to find β_j (the weight of the j^{th} possible hub) by using multi criteria decision making model.

Cities with land border were taken as demand nodes (Table 1). Existing airports in Turkey were chosen (Fig. 2) while possible hubs were being determined (Table 2). Therefore, the installation cost of the hubs was not considered.

Fig. 2. Airports

Table 1. Demand nodes

NUMBER OF DEMAND NODES	DEMAND NODES
1	HATAY
2	KİLİS
3	G.ANTEP
4	URFA
5	MARDİN
6	ŞIRNAK
7	HAKKARİ – IRAN
8	HAKKARİ - IRAQ
9	VAN
10	AĞRI
11	IĞDIR
12	KARS
13	ARDAHAN
14	ARTVİN
15	EDİRNE
16	KIRIKKALE

Table 2. Potential hubs

NUMBER OF HUBS	POTENTIAL HUBS
1	ADANA
2	ADIYAMAN
3	AĞRI
4	ANKARA
5	ANTALYA
6	BALIKESİR
7	ÇANAKKALE
8	DENİZLİ
9	DİYARBAKIR
10	ELAZIĞ
11	ERZİNCAN
12	ERZURUM
13	GAZİANTEP
14	ISPARTA
15	İSTANBUL
16	İZMİR
17	KARS
18	KAYSERİ
19	KONYA
20	MALATYA
21	MARDİN
22	KAHRAMANMARAŞ
23	MUĞLA
24	MUŞ
25	NEVŞEHİR
26	SAMSUN
27	SİİRT
28	SİNOP
29	SİVAS
30	TEKİRDAĞ
31	TRABZON
32	ŞANLIURFA
33	UŞAK
34	VAN

Proposed mathematical model is built by assumptions which determine which possible hub(s) would be appropriate to be used.

B. Assumptions

- The aircraft altitude is constant.
- There is always a sight available.
- The model of all the aircrafts is same.
- The speed of the aircrafts is constant (220km/h).
- The speed of the aircrafts during the observation operations is 110 km/h.
- Observation diameter is 15 km.
- The communication range of the aircrafts is 300 km.

- Aircrafts turn back to their hubs.
- Duties are performed in good weather conditions.
- There is no threat for the aircrafts.
- Considered hubs are controled by General Directorate of State Airports Authority of Turkey.

C. Parameters

H ; $\{1,...,m\}$ is the sequence of the main possible hubs

N ; $\{1,...,n\}$ is the sequence of the demand nodes.

i ; demand node $i = 1,2,...,16$

j ; possible hub $j = 1,2,...,34$

β_j ; the weight of the j^{th} possible hub

d_{ij} ; distance between i^{th} demand node and j^{th} hub

S ; the maximum range of the aircraft

p ; total number of the hubs (p<=the number of the demand nodes)

$$a_{ij} = \begin{cases} 1, d_{ij} < S \\ 0, otherwise \end{cases}$$

g_i; the number of the illegal transportation at the i^{th} node in (Jan-June 2009)

o_i ; the number of the illegal incidents at the i^{th} node in (Jan-June 2009)

$g_i + o_i$; the demand of the i^{th} node

D. Decision Variables

$$Yj = \begin{cases} 1, \text{if possible hub j is selected as a hub} \\ 0, \text{otherwise} \end{cases}$$

$$Xij = \begin{cases} 1, \text{if node i is connected to the hub j} \\ 0, \text{otherwise} \end{cases}$$

E. Model

$$\max Z = \sum \beta_j * Y_j + \sum_i^n \sum_j^m (g_i + o_i) * X_{ij} \tag{1}$$

$$\sum_j^m X_{ij} = 1 \qquad \forall i \in N \tag{2}$$

$$X_{ij} \leq Y_j \qquad \forall i \in N \qquad \forall j \in H \tag{3}$$

$$\sum_j^m Y_j = p \tag{4}$$

$$X_{ij} \leq a_{ij} \qquad \forall i \in N \qquad \forall j \in H \tag{5}$$

$$X_{ij} \in \{0,1\} \qquad \forall i \in N \qquad \forall j \in H \tag{6}$$

Equation (1) defines the aim function. **It** decides about which hub will be activated for maximum benefit using weighting parameter. Equation (2) assigns a demand node for each hub using (3). The maximum number of the active hubs is limited by p with (4). Equation (5) does the hub selection and assignment using the maximum range of the aircraft.

F. Results

The GAMS/CPLEX solver is used to select the hubs to be activated among the possible hubs.

Model is solved for each p value individually while $p = 1,2,...,16$. Current airports are defined as possible hubs. Therefore, the cost of adding a new hub is not considered. According to the results of our model, Tekirdag, Trabzon, Gaziantep, and Van airports are decided to be used for UAV system (Table 3).

Table 3. Results

Number of Hubs	Recommended Hubs	Related Demand Nodes	Aim Function
p=4	(13) ANTEP	1,2,3,4	650,1770
	(30) TEKIRDAG	15,16	
	31) TRABZON	13,14	
	(34) VAN	5,6,7,8,9,10,11,12	

The reasons of recommending Tekirdağ as a hub instead of Istanbul are to be close to the land border, to have a current airport, and to have low traffic. The weight factors of the other hubs were effective on the decision. The weight factors are calculated using the following criteria:

i) whether the airport belongs to the military or not,
ii) whether there is threat in the area or not,
iii) whether the city has a land border or not,
iv) whether the airport has low air traffic or not, and
v) the duration that the airport is in service.

Our results show that, the weight factors that are calculated using many criteria are very useful.

IV. CONCLUSION

The problem of the single allocation p-hub median is addressed to monitor the movement at Turkey's land borders by using UAV. The range of the UAV is used to determine the best location to install the UAV system. In future work, the cost and the altitude of the UAV may be added to the model.

REFERENCES

[1] A.T. Ernst, H. Hamacher, H. Jiang, M. Krishnamoorthy, and G. Woeginger, "Uncapacitated single and multiple allocation *p*-hub center problems", *Computers and Operations Research*, 36, pp. 2230-2241, 2009.

[2] S. Alamur and B.Y. Kara, "Network hub location problems", *European Journal of Operational Research*, 190, pp.1-21, 2007.

[3] J. Kratica, Z. Stanimirovic, D. Tosic, and V. Filipovic, "Two genetic algorithms for solving the uncapacitated single allocation p-hub median problem", *European Journal of Operational Research*, 182, pp.15-28, 2007.

[4] J.F. Campbell, "Integer programming formulations of discrete hub location problem", *European Journal of Operational Research*, 72, pp. 387-405, 1994.

[5] J.F. Campbell, "Hub location for time definite transportation", *Computers & Operations Research*, vol.36, pp.3107-3116, 2009.

[6] S. Elmastaş, "Hub location problem for air ground transportation systems with time restrictions", PhD Thesis, Bilkent University, 2006.

[7] J. Sohn and S. Park, "A linear program for the two-hub location problem", *European Journal of Operational Research*, 100, pp.617-622, 1997.

[8] J. Ebery, "Solving large single allocation p-hub problems with two or three hubs", *European Journal of Operational Research"*, 128, pp.447-458, 2001.

[9] H. Yaman, B.Y. Kara and T. Barbaros, "The latest arrival hub location problem for cargo delivery systems with stopovers", Transportation Research, Part B(43), pp. 643–658, 2009.

[10] A. Özger and H. Oktal, "***I***nteger model approach for uncapacitated multiple allocation *p*-hub median problem in air cargo transportation", *Havacılık Ve Uzay Teknolojileri Dergisi*, 4, pp.47-60, 2009.

[11] www.stratejikboyut.com.article_detail.php (Cafer Serinkan)

[12] http://www.fas.org/irp/program/collect/uav.htm

[13] http://www.globalsecurity.org/intell/ systems/ uav-intro.htm

LPT to RS232 communication converter

Grofu Florin “Constantin Brancusi” University – Tg-Jiu
Cercel Constantin “Constantin Brancusi” University – Tg-Jiu

Abstract: **The base idea of this paper came as a result of using of some data acquisition systems controlled by an application made in LabWindows CVI. The communication between PC and the acquisition systems is made through RS232 interface. The main problem appeared is the impossibility to use the acquisition system for fast bidirectional data transfer using a RS232 protocol. As solution for this problem we propose a LPT-RS232 communication converter.**

Keywords: Acquisition System; Communication; RS232 interface; Converter

I. Introduction

The achievement of data acquisition systems by an automation engineer is an easy thing to do. This can be made by connecting a microcontroller to an AD or DA converter, depends on application. For communication of this system with a software application the easiest way is to use the RS232 interface. The schematic diagram is presented in figure 1.

For hardware specialist is easy to implement for the acquisition system a communication protocol which can ensure the data transfer in optimal condition without blocking the system or misunderstanding the data.

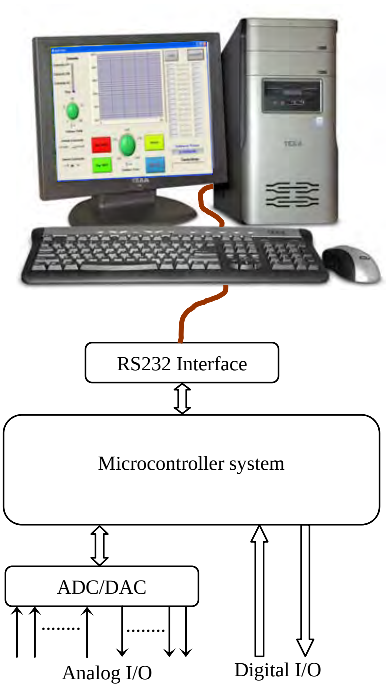


Fig. 1 Data acquisition system

In case of using of a 8051 family microcontroller there is the possibility to use of multi processor communication.

In these modes, 9 data bits are received, followed by a stop bit. The ninth bit goes into RB8. Then comes a stop bit. The port can be programmed such that when the stop bit is received, the serial port interrupt is activated only if RB8 = 1. This feature is enabled by setting bit SM2 in SCON. When the master processor must transmit a block of data to one of several slaves, it first sends out an address byte that identifies the target slave. An address byte differs from a data byte in that the 9th bit is 1 in an address byte and 0 in a data byte. With SM2 = 1, no slave is interrupted by a data byte. An address byte, however, interrupts all slaves, so that each slave can examine the received byte and see if it is being addressed. The addressed slave clears its SM2 bit and prepares to receive the data bytes that follow. The slaves that are not addressed set their SM2 bits and ignore the data bytes.[4]

In this data acquisition system data we use the follow structure of data package, presented in table 1.

This structure allows to connect more acquisition systems to the same software application. If the application sends one data package all the systems receive on the interrupt first byte, which mean the address of destination module using multiprocessor mode. Only the module with this address goes in data receive mode, the rest of system ignore the data that follows. On the complete receive of package it check the CRC as an additional correction.[1]

Table 1

Octet 1	Address byte	address byte
Octet 2	Nr. bytes	data bytes
Octet 3	Actions	data bytes
Octet 4	Data	data bytes
....	....	....
Octet n	Data	data bytes
Octet n+1	CRC	data bytes

II. Software aplication

The software application for connecting with the system it is realized in LabWindows CVI.[7] LabWindows/CVI is an integrated American National Standards Institute (ANSI) C environment (C is a computer programming language) developed by National Instruments Corporation and designed primarily for engineers and scientists creating virtual instrumentation applications. LabWindows/CVI helps you leverage the power of the computer to create flexible, reusable, and inexpensive measurement applications that outperform traditional test and measurement methods.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_42, © Springer Science+Business Media B.V. 2010

LabWindows/CVI is a programming environment that has been widely adopted throughout industry, academia, and research labs as the standard for data acquisition and instrument control software. It is a powerful and flexible instrumentation software system that embraces the traditional programming methodologies and enables you to create structured code, and it features an easy-to-use programming environment.[8]

LabWindows/CVI includes all the tools necessary for data analysis and presentation of the results on the Graphical User Interface (GUI) on your computer screen. It comes with a complete set of integrated input/output (I/O) libraries, analysis routines, and user interface tools that aid you in rapid application development, reducing the time that would be required by other conventional programming environments.[6]

LabWindows/CVIs forte lies in building virtual instrumentation systems with General Purpose Interface Bus (GPIB), VME (Versa-Modular Eurocard) extensions for Instrumentation (VXI), PCI (Peripheral Component Interconnect) extensions for instrumentation (PXI), serial interface communication (RS-232), Transmission Control Protocol/Internet Protocol (TCP/IP) based devices with plug-in data acquisition (DAQ) boards without spending too much effort to create the applications. It combines an interactive, easy-to-use development approach with the programming power and flexibility of compiled ANSI C code.

The full potential of LabWindows/CVI is used to automate test systems, bench top experiments, DAQ monitoring projects, verification tests and measurements, process monitoring, and controlling systems.

III. LPT-RS232 CONVERTER

In case we use the RS232 communication between PC and data acquisition system up to 10 times per second the application is easy to use.

If we need to increase the speed of communication the problems begin to appear and only a very good LabWindows programmer can resolve them. This problem appears as effect of serial port accessing mode through the existing functions. The communication is delayed because the ninth bit (parity bit) must be different set for a command byte or a data byte.

From experiments we observed that delays occur when accessing the change configuration functions of the serial port or on change direction of transfer. Thus if the complete communication means transmission and reception of a package of 6 bytes of data at a transfer rate of 9600 bits / sec, theoretically should last up to 15 ms, practically takes about 100 to 150 ms depending on the architecture and performance of computer system used.

This leads to a decrease of almost 10 times of the effective speed of communication resulting the possibility of a bidirectional data transfer up to 10 times per second.

An alternative solution to serial communication is the parallel communication that provides high speed communication but on short distance. If there is already exist serial communication equipment then must be implemented a parallel to serial communication converter. This converter simply takes parallel data package and then sends it serial to acquisition system. It also takes serial data package from the acquisition system and forwards the data parallel to the application.[9]

Thus, all changes of state of serial communication port is transferred to the parallel to serial conversion module that carried out in real time. The structure of such a communication system is shown in figure 2.

The converter is made bidirectional, meaning it can transmit and also receive data, so the information will flow not only from computer to peripheral, but also from the computer peripheral.

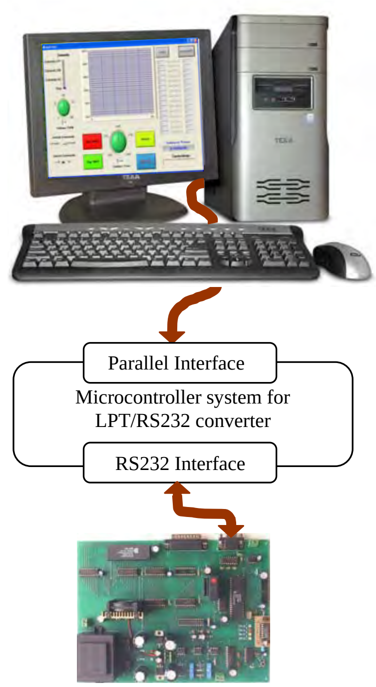


Fig. 2. Using LPT/RS232 converter

Knowing the communication protocol for serial transfer, how to transfer data using PC parallel port and the structure and functioning of a 89S52 Intel microcontroller we designed and achieve a parallel to RS232 communication converter.

Functional diagram of the converter is shown in figure 3:

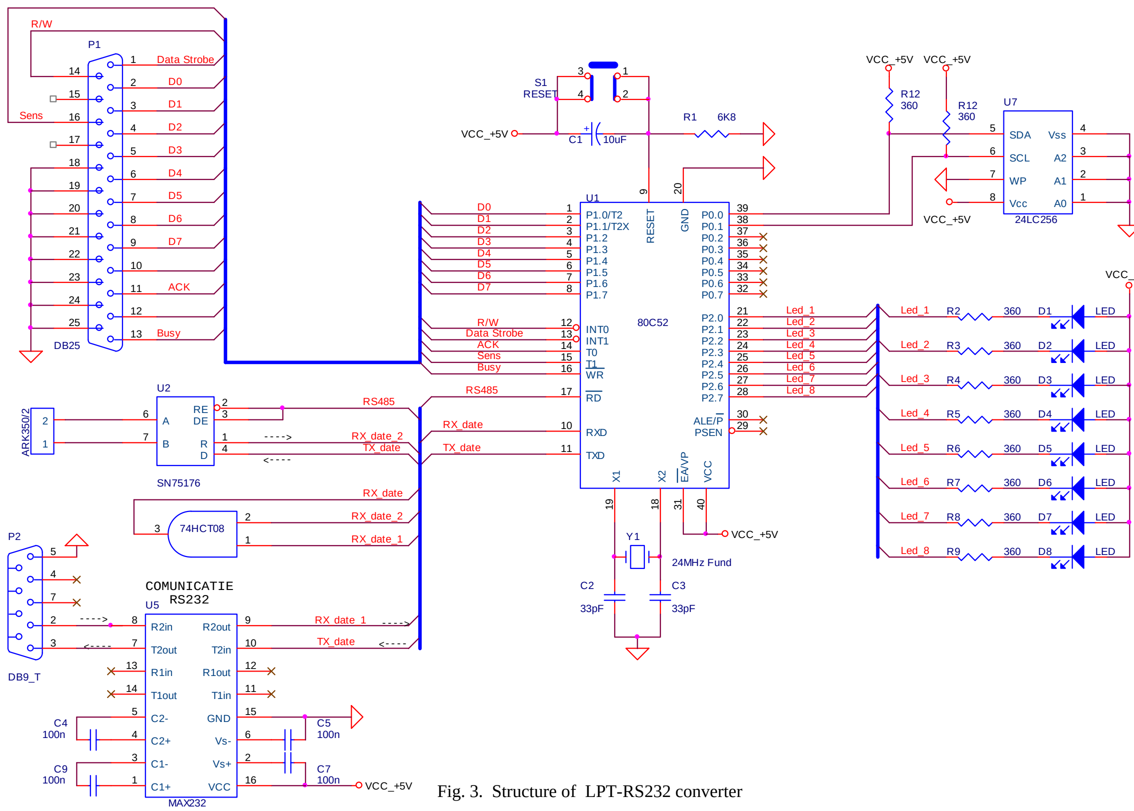


Fig. 3. Structure of LPT-RS232 converter

In structure of converter we see an EEPROM memory which is necessary to use the converter in more complex application were we transfer more then 128 bytes. The EEPROM memory is used also for back-up some setting of module.

To explain better the advantage of using this type of converter we achieve a software application that we connect three identical acquisition systems to the PC in this way:

- one directly to PC serial port (COM1);
- one through an USB to serial converter;
- one through the proposed parallel to serial converter.

The application, with the graphical interface presented in figure 4, send sequential by the three acquisition systems a 6 bytes data package. In case of correct communication the acquisition systems replay also a 6 bytes data package.[1]

As result of our experiments we obtain the follows communication times as in table 2.

Table 2

Nr. Crt.	DAS communication	Time communication
1.	RS232 (COM1)	402,7 ms
2.	USB-RS232 converter	372,3 ms
3.	LPT-RS232 converter	14,6 ms

CONCLUSION

From performed analyze result that in case of acquisition system with serial communication and where is necessary to increase the effective communication speed we need to contact a programmer which may realize a software module which can be executed faster (in LabWindows), module which depend on the hardware structure of computer.

A simply way is to use the proposal parallel to serial converter which ensure the increase of effective communication speed indifferently of computer architecture considering that to send and receive an 6 byte data package takes on parallel port approximately 1ms.

In figure 4 it is presented the interface of the software application which can demonstrate the advantage of proposed communication system. The software application has the possibility to communicate with three identical acquisition devices connected with PC by different ways: one with direct connection to PC serial port (RS232), one using an USB to RS232 converter, and one using the proposed LPT to RS232 converter.

As we can see in the figure 4, if we consider the values displayed in the Time Value boxes, the time for communication corresponding to the proposed converter is at least ten times smaller than others. So, the real speed of communication is increased with at least ten times. We consider this time the time necessary to send and receive a data package.

The LPT to RS232 converter is very useful in situations that the data packages sent and received is not so big (under 128 bytes), case occurred in monitoring and control applications for industrial process.

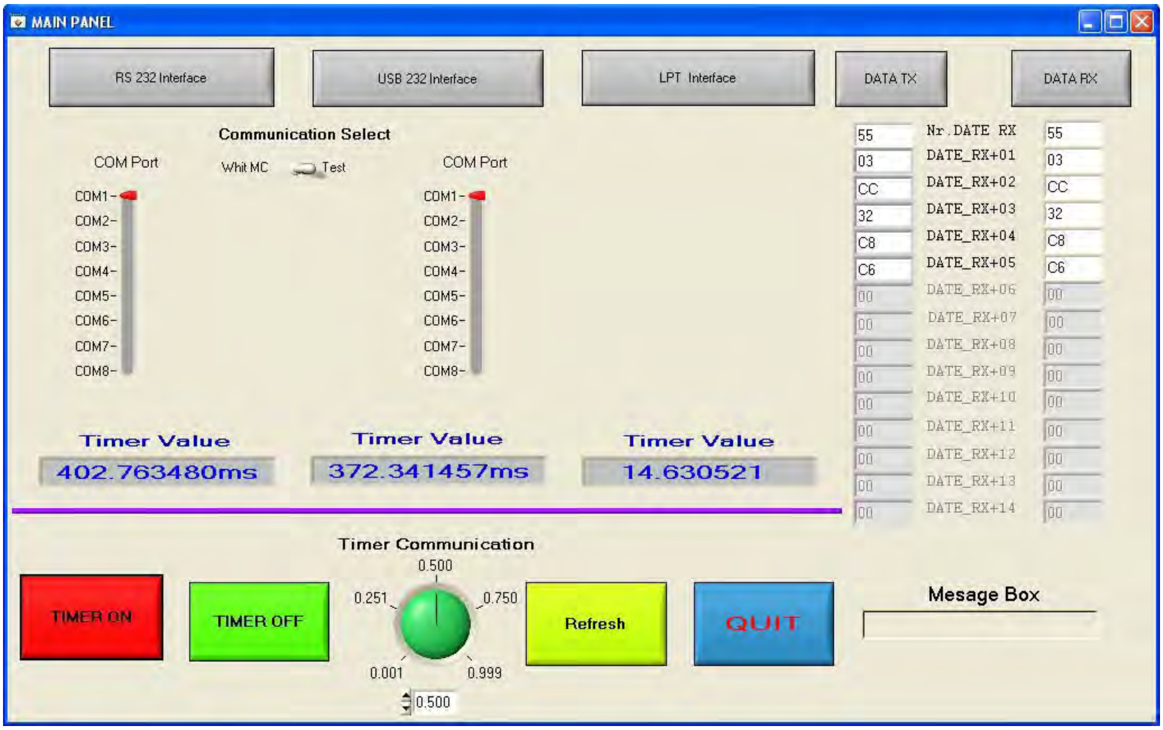


Fig. 4. User interface

REFERENCES

[1]. Cercel Constantin**,** Grofu Florin, *USB to Parallel communication interface***,** Annals of „Constantin Brancusi" University of Tirgu Jiu, The XIII-th edition of the international scientific conference of the Engineering Faculty, November 07-08, 2008 , ISSN 1844-4856

[2]. Porat B., *A Course in Digital Signal Processings*, John Wiley & Sons, Inc., New York, 1997.

[3]. Thomas B. Cho, *Circuit Techniques for Low Power, High Speed Pipelined A/D chapter 2*, University of California, Berkeley 2000

[4]. Dallas Semiconductor, *Ultra-High-Speed Flash Microcontroller User's Guide* , http://www.maxim-ic.com/ 2004

[5]. Maxim-IC, *App Note 604: Fast Memory Transfers with the Ultra High-Speed Microcontroller - DS89C420,*

[6]. www.ni.com/pdf/manuals/373551c.pdf

[7]. Vektrex Application Note: 11082002, Using IVI-COM Drivers From LabWindows/CVI

[8]. Advanced Topics in LabWindows/CVI by Khalid, Shahid F

[9]. Grofu Florin, Popescu Luminita Georgeta, Cercel Constantin *Acquisition system for monitoring vibrations*, CISSE2007 – Novel Algorithms and Techniques in Telecommunications, Automation and Industrial Electronics, Springer, 2008, ISBN: 978-1-4020-8736-3

[10]. Runceanu Adrian, *Fragmentation in distributed databases*, CISSE2007 - Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, Springer, 2008, ISBN: 978-1-4020-8734-9

Simulation Model of Coremaking Plant

Stefan Belev[a], Miroslav Penchev[b],
[a,b]University of Rousse, 8 Studentska str., 7017 Rousse, Bulgaria

Abstract

The paper presents simulation model of coremaking plant which gives a possibility for optimal utilization of working machines. The aim of modeling is investigation of coremaking equipment idle time in dependence of a section production plan within a working day. The model is realized like simulation program which is built in the system for operative management (SCADA) at one of coremaking plants in Vidima Ltd. Sevlievo, Bulgaria.

Keywords: coremaking plant, simulation model

1. Introduction

The main method applied for production of casting cores for needs of non-ferrous metallurgy and mostly for producers of sanitary fittings is the hotbox method. In this coremaking technique, alkali free dry silica sand is mixed with latent (heat activated) hardener and liquid binder (Phenolic resole or Urea-modified furan resin). The ready made moist mixture for casting cores contains about 1.4 – 2.0 % resin and can be used within 4-6 hours, depending on the type of the used resin. The mixture is transferred to an automatic core-blowing machine where the core is moulded and cured at 180-250°C in a previously heated corebox, Fig. 1.

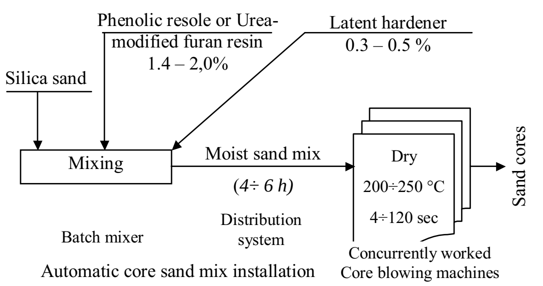


Fig. 1. Coremaking plant - hotbox processes.

Modern coremaking plants are equipped with fully automated sand mixing equipments including a batch mixer and a core sand mixture distribution system. The core sand mixture distribution system is designed and adjusted to the particular coremaking plant depending on the specific plant architecture and the customer's requirements. Usually the distribution systems which transfer the ready core sand mixture from the mixer to the core-blowing machines is based on railway hopper cars or special suspended monorail cars, technological scheme – "mobile container". In the railway hopper cars technological scheme the variant "mobile mixer" is also possible.

Both technological schemes (mobile container and mobile mixer) are based on *servicing a group of concurrently operating machines by one mixer and one transport device.* The main advantage of these schemes is the economy, because one mixer and one transport device serve all working machines. The potential of delaying the service, and appearing of the idle time could be pointed as a disadvantage, because the queue is formed when several machines must be served at the same time. It should be noted that in most cases the separate core-blowing machines produce different products which increases the risk of idle time caused by lack of servicing because of the different technological times and the different sand core mixture consumption.

The task of the developed model is *to show the idle times caused by lack of servicing* in quantitative (as idle times for the separate machines) and qualitative (as moments of occurrence of the idle times) aspect *depending on the daily job of the plant.*

2. Conceptual model of coremaking plant

The investigating plant consists of 10 linearly situated core-blowing machines, applying the **hotbox** method. The mixing equipment, connected to the coremaking plant is of the mobile container type, Fig. 2.

2.1. Model hierarchical structure

A hierarchical model of the separate components, derived through functional decomposition, may represent the coremaking plant.

The plant produces about 300 types of casting cores to satisfy the needs of a foundry for sanitary fittings. For the purposes of this research, the following may be pointed as characteristics of the products: the time for the technological cycle T_{C_J} and the characteristics for the consumption of sand

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_43, © Springer Science+Business Media B.V. 2010

core mixture (the mass of the respective product m_J and the number of products, made within one technological cycle n_J). The index "**J**" designates the type of the produced item:

$$\begin{cases} Z = \{z_1, z_2, \dots, z_J\}; J = 1,2,\dots,300 \\ z_j = z(Tc_j, m_j, n_j), \forall z_J \in Z \end{cases} \quad (1)$$

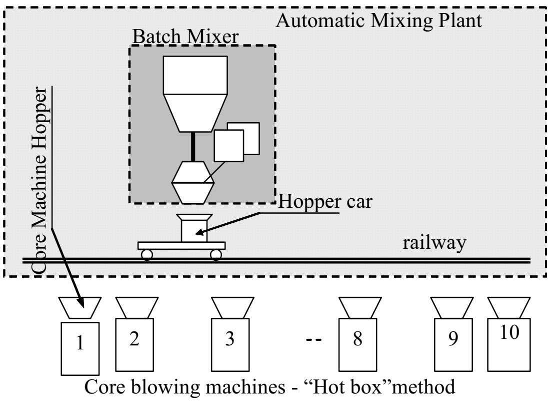


Fig. 2. Coremaking plant – architectural diagram.

The daily job may be represented as a section of the multitude of the machines (2) and the multitude of the produced items (3). The index **I** designate the position of the core-blowing machine in the technological scheme of the plant.

$$Mh = \{mh_1, mh_2, \dots, mh_I\}; I = 1 \div 10 \quad (2)$$

$$MP_E = Z \cap Mh \quad (3)$$

Each of the **core-blowing machines** in the plant is equipped with a hopper for sand core mixture with adjustable minimum level r_I, (in % of the hopper's content). When the mixture level falls below the preset minimum, the machine transmits request for serving to the mixing equipment[1]:

$$R_I = R(I, d, t), \quad (4)$$

where:

- I is the location of the core-blowing machine which has transmitted a request;
- d - the number of the requested (by the respective machine) doses of sand core mixture (**d = 1** or **d = 2**);
- t - the moment of originating of the request, as a moment of occurrence of a characteristic event, changing the condition of the observed system.

In this case d and I are the request parameters.

As for the aims of this research, it could be considered that a united flow of requests is received by the mixing equipment:

$$R = \bigcup_{I=1}^{10} R_I(I, d, t) \quad (5)$$

The mixing equipment as an *algorithm for servicing the requests* may be considered as a queuing system with standard **(FIFO)** discipline of servicing. The equipment parameters are determined and can be presented by the set **Mix**:

$$Mix = \{d, M, T_{MX}, Ttr_I, Tul\}, \quad (6)$$

where:

- d is the number of the currently requested doses of sand core mixture;
- M - the mass of one dose of sand core mixture;
- T_{MX} - the time necessary for the preparation of one dose of sand core mixture;
- Ttr_I - the time necessary for conveying the mixture to the hopper of the core shooting machine, which has sent the request;
- Tul - the time necessary for discharging the mixture into the machine's hopper.

Some special features of the mixing equipment follow from the plant's architectural scheme, shown in Fig. 2. They are:

- the service device has a constant component – the working time of the mixer $T_{MIX} = dT_{MX}$ defined by parameter d;
- the service device has a different time for conveyance ($T_{TR_I}, I = 1 \div 10$), depending on the route, defined by parameter I.

In the conceptual model of coremaking plant (see Fig. 3) these special features are described by denoting the events:

- Elimination of the request after loading the core shooting machine **- Departure (I,t)**.
- Readiness of the transporting device - **Car Ready (t).**

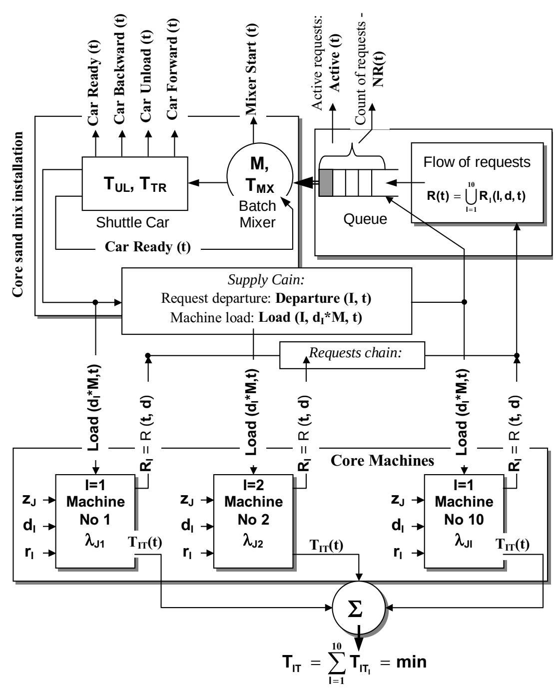


Fig. 3. Conceptual model of coremaking plant.

In our case **the subject of investigation** is a casual flow of events: moments of turning off & on and the duration of intervals

[1] The mixture consumption model in the hopper is reviewed in following section.

$\mathbf{T_{IT}}$ during which the separate machines do not work because of lack of sand core mixture (Fig. 3 and Fig. 4).

The aim of the investigation (in order to enhance the effectivity of loading of the core-blowing machines) may be defined as: *minimizing the idle time caused by lack of servicing of the machines in the plant* (7), by changing the location of the product of the current daily job in the technological scheme of the plant.

$$T_{IT} = \sum_{I=1}^{10} T_{IT_I} = \min. \qquad (7)$$

2.2. Core Blowing Machine model

A model of consumption of the mixture may represent the moment of sending the request and an idle time interval that is caused by lack of sand mixture, Fig. 4.

The following denotations are used in the Figure:

L – current hopper level, %;
r – request level, %:

$$\text{Request} = \begin{vmatrix} 0 & L > r \\ 1 & 0 \le L \le r \end{vmatrix} \qquad (8)$$

$\mathbf{T_C}$ – core blowing machine cycle, sec;
$\mathbf{T_{AT}}$ – arrival time interval, sec;
$\mathbf{T_{WT}}$ – waiting time interval, sec;
$\mathbf{T_{IT}}$, – idle time interval, machine does not work, sec.

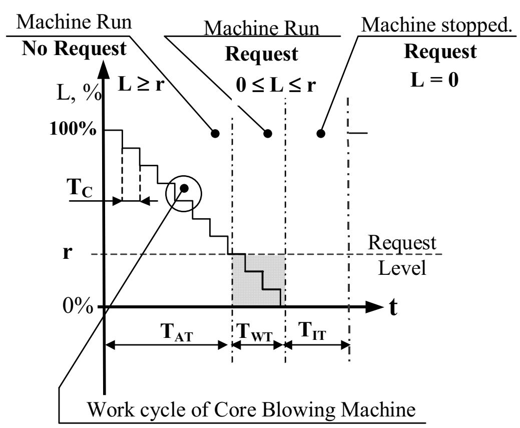


Fig. 4. Core blowing machine – hopper level.

Taking into consideration the discrete character of the machine cycle and the casual character of the interval $\mathbf{T_{WT}}$ for the mixture level in the hopper of machine **I**, producing item **J** at the moment of supplying, it could be written down:

$$\begin{vmatrix} L_{I,J}(t, k_J T_{C_J}) = L_{I,J}(t, k_J T_{C_J}) + 100 * d_I, \% \\ k_J = 0 \end{vmatrix}, \qquad (9)$$

and for the level after each cycle:

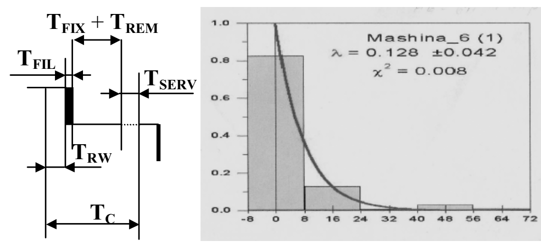
$$\begin{vmatrix} L_{I,J}(t, k_J T_{C_J}) = L_{I,J}(t, (k_J - 1)T_{C_J} - 100 \dfrac{m_J n_J}{M}, \% \\ L_{I,J}(t, k_J T_{C_J}) \ge 0 \\ k_J = k_J + 1 \end{vmatrix}, \qquad (10)$$

where: **t** is the current time of the simulation experiment, and $\mathbf{k_J}$ – the current number of the machine cycle from the moment of loading the machine.

The individual technological cycle may be described by equation (11), Fig. 5a:

$$T_C = T_{RW} + T_{FIL} + T_{FIX} + T_{REM} + T_{SERV} = T_{MhC} + T_{SERV} \qquad (11)$$

where: $\mathbf{T_{RW}}$ is the time for moving the operating organs from initial to operating position; $\mathbf{T_{FIL}}$ – the time for filling and congesting the mould; $\mathbf{T_{FIX}}$ – the time for curing the core into the mould; $\mathbf{T_{REM}}$ – the time for opening the mould, removing the core and returning the operating organs to initial position; $\mathbf{T_{SERV}}$ – time for cleaning and lubricating the corebox.



a) Cycle Time **b)** Service Time Model

Fig. 5. Core blowing machine – technological cycle.

The time intervals $\mathbf{T_{RW}}$, $\mathbf{T_{FIL}}$, $\mathbf{T_{FIX}}$ and $\mathbf{T_{REM}}$ are functions of the machine settings for the particular product **J**. And the operation of cleaning and lubricating the corebox is performed manually by the operator and is of probability nature. Observation of the work of machine 6 with product: faucet "bath body SL – A" is shown in Fig. 5b. The realized observations and statistic investigations show that the time for service may be described by an exponential distribution which parameter is defined by the average time.

3. Model implementation

On a system level the simulation model assumes description of the functioning of the simulated system using the concepts activities, events and processes by means of a definite programming language. The methods known in practice for organizing the computing process in the simulation programs (activity scanning, event schedule and processes interaction) are striving to formalize and solve the problem with the concurrent processes, common to all computer simulators. In this particular case due to the elementary organization of the simulation machine and the relatively easy realization by means of a universal programming language, for realization of the simulator the "**activity scanning"** method is chosen.

3.1. Model formalization

Definition 1: The dynamic model of a simple object may be represented as an arranged five, [1]:

$$\mathbf{DM = (S; E_{INP}; E_{OUT}; SC; g)}, \qquad (12)$$

where **S** is the endmost set of states of the object;
$\mathbf{E_{INP}}$ – the set of the input events of the object;
$\mathbf{E_{OUT}}$ – the set of the output events of the object;

$\mathbf{g : S \times E_{INP} \rightarrow S}$ - transition function.

SC – the set of conditions, which may be defined as a subset of the events, admissible for the separate states $\mathbf{SC : \forall s \in S; E^{S}_{INP} \subset E_{INP}}$.

This model of the states and transitions is interpreted in the following manner. If the object is in state $\mathbf{S_I}$ and event $\mathbf{e \in E_{INP}}$ occurs, which is admissible for that state, then the object transits in a new state defined by the function $\mathbf{g : S \times E_{INP} \rightarrow S}$.

The visual formalism of that interpretation is "State Transition Diagram", well known from the theory for the finite state machine and widely used in automation.

In this paper the diagram specification defined by IEC 61131 [2] is used. It is the base of many programming languages for controlling devices. *The diagram is a directed graph whose points (denoted by a circle) represent the object's states* and the directed line denotes the transition to the next state (Fig. 6). In a rectangle to the right of the state the actions performed by the object in that state are denoted. In a random moment of time the object may be only in one state – *active state*.

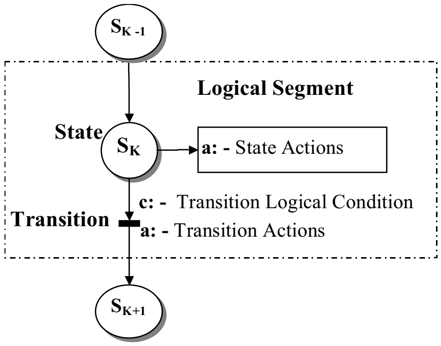


Fig. 6. State transition diagram – definitions.

Definition 2. The transition is instantaneous and may be activated if the preceding state is active and corresponding logical condition is executed. <u>*The activation of the transition*</u> leads to activation of the next state and deactivation of the preceding state. *One or several single actions may be associated to the transition.*

Tow special states are foreseen: <u>*initial state*</u> (initialization), in which all the rest of the states are brought to zero, and "<u>*blocking*</u>" state (stop) which blocks the work of the object. The object can leave that state only at initialization.

Since the specification does not provide for synchronization between the actions of the separate objects and measuring the lifetime of a separate state according to the model time, the following additional rules are accepted for the purposes of the simulation:

1. Each initial for the object event is registered by means of setting of a synchronizing signal.
2. After occurrence of an incoming event the object nullifies the signal, which has initiated the event.
3. The object settles in initial state only through external signals, by means of transition to initial state.
4. The lifetime of the separate state is associated to the model time as an action in the transition preceding the state: $\mathbf{t_{SI} = t + T_{SP\_SI}}$. Here **t** is the current moment of the model time in which the transition is activated, and $\mathbf{T_{SP\_SI}}$ is the preset and/or calculated time of life in state $\mathbf{S_I}$. (Computation of the times is performed beforehand).

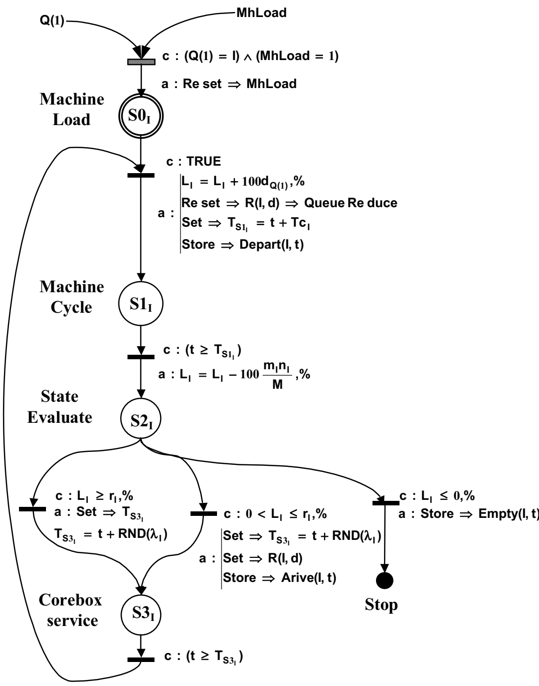


Fig. 7. State transition diagram of core machine model.

Figure 7 represents the atomic model of the core-blowing machine which is developed according to specification IEC 61131 and the rules considered above. The overall model of the core plant reflects the synchronization between the associated elementary models of the objects participating in the modeled production system. They are the group of 10 concurrently operating core-blowing machines; mixing equipment including batch mixer and hopper car; a queue with a standard FIFO discipline and a database for the simulation experiment.

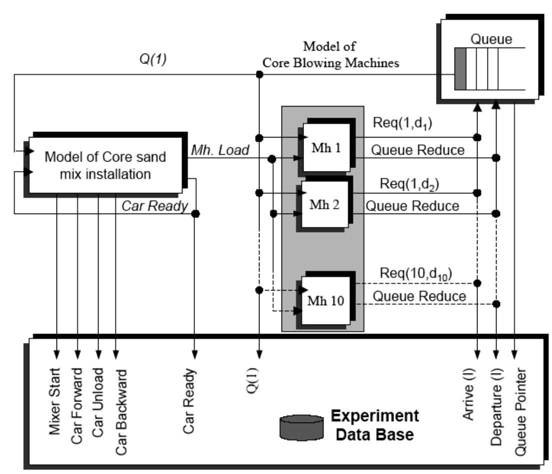


Fig. 8. Coupled model of coremaking plant.

ID	Time	Requests	Mash No	Arive	Mash Set Req	Departure	Empty	Mixer Start	Car Forward	Car Unload	Car Backward	Car Ready
171	9330	0	0	0	0	1	0	0	0	0	0	0
172	9330	0	0	0	0	0	0	0	0	0	0	1
173	9500	1	6	1	6	0	0	0	0	0	0	0
174	9500	1	6	0	0	0	0	1	0	0	0	0
175	9570	2	6	1	5	0	0	0	0	0	0	0
176	9615	2	6	0							0	0
177	9620	3	6	1							0	0
178	9653	3	6	0							0	0
179	9660	3	6	0							0	0
180	9668	3	6	0	0	0	0	0	0	0	1	0
181	9669	2	6	0	0	1	0	0	0	0	0	0
182	9713	2	5	0	0	0	0	0	0	0	0	1
183	9714	2	5	0	0	0	0	1	0	0	0	0
184	9829	2	5	0	0	0	0	1	1	0	0	0
185	9866	2	5	0	0	0	0	0	0	1	0	0
186	9874	2	5	0	0	0						0
187	9875	1	5	0	0	1						0
188	9911	1	4	0	0	0						1
189	9912	1	4	0	0	0						0
190	9932	1	4	0	4	0	1	0	0	0	0	0
191	10027	1	4	0	0	0	0	1	1	0	0	0

Event No 173: Request from Core Machine 6
Modeling Time: 9500 sec
Request in Queue:1

Event No 181: Core Machine 5 - Full
Modeling Time: 9669 sec
Request in Queue:2

Fig. 9. Result of modeling experiment.

The chosen model for organizing the computation process of the simulator (organization of the simulation machine) does not reflect the lifetime of the separate states of the objects participating in the simulation system. This problem is solved by means of registering the events which lead to activation and deactivation of the states concerning the user, into a database. For this purpose, the respective events are declared outgoing and synchronizing signals are sent to the database, which are processed according to the above reviewed rules. At the consequent processing of the database the lifetime of the respective state is calculated as the difference between the model times of deactivating and activating events, Fig. 8.

The simulator was realized using **C++** and was built in the operating control system of the core plant (SCADA package WinCC 6.0) in VIDIMA JSC, Sevlievo, Bulgaria. The widely available data base software MS Access was used, which is also supported by the SCADA package. The results of modeling experiment are shown in Fig. 9.

4. Conclusion

The SCADA systems usually do not include a simulator of the controlled system, which could be used for supporting short-term management resolutions. The developed simulator affords the opportunity for measuring the time, during which the simulated plant remains in the separate states, as well as a detailed tracing of the moments of occurrence of the events and may be used as an instrument by the plant management staff.

It could be considered that the method applied for the simulator development is applicable to the practical tasks for measuring the time of the separate activities of relatively simple industrial systems with random events, for which a detailed tracing of the moments of occurrence of the individual events is necessary.

References

[1] A. M.Law, W. D. Kelton: *Simulation Modeling and Analysis*, Mc. Graw-Hill, 2000.

[2] W. Lewis. *Programming Industrial Control Systems Using IEC 1131-3, volume 50 of IEE Control Engineering Series*. The Institution of Electrical Engineers, London, UK, 1996.

Depth Calculation using Computer Vision and Sift

Edson CAVALCANTI Neto*, Haroldo Cabral MAYA *, Antonio Themoteo VARELA, *
Mayara Machado CARVALHO*, Auzuir Ripardo de ALEXANDRIA *
*IFCE - Instituto Federal de Educação, Ciência e Tecnologia do Ceará
Av. Treze de Maio, 2081 - Benfica
Fortaleza, Ceará, Brazil, +55 (85) 3307-3666
Email: edsoncavalcanti@hotmail.com; haroldomaya@hotmail.com; themoteo@ifce.edu.br;
mayaracarvalho@hotmail.com; auzuir@ifce.edu.br

***Abstract*—Robotics has been developed in a high level in the past few years, especially in mobile Robotics. Its main contribution is shown in human entertainment and performance in dangerous works. However, those are not just the only applications they perform. There is a wide field to be developed. Distance calculation is something really important for localization and navigation of mobile robots. The classic way to calculate distances are using sonar and laser radar. But, the use of computer vision in stereo, based on two cameras, is widespread in Robotics. Therefore, algorithms, to detect new points, allows the creation of new methods to execute stereo vision. In those algorithms there is one that has been used, the SIFT (scale-invariant feature transform). It may be used with classic algorithms of stereo vision in the Computer Vision 3D system development. This project's objective is to develop a distance acquisition system using computer vision in stereo with the help of SIFT method. The results are very close with what we were expecting. The next steps to be done are the execution of quantitative tests to determinate the focal distance of the utilized cameras and to calibrate the system.**

I. Introduction

A typical mobile robot is constituted by subsystems such as: perception, localization, navigation and movement control [1]. The perception system is responsible to acquire data from the space where the robot is placed. This data, after being stored, is processed to be used to generate information on the location where the robot is placed, through its location system. In possession of its location, the robot is able to generate the environment map where it is situated. After it may compare its local position with a more comprehensive map generated or delivered by its developer or user [2]. The ability to navigate using the map and the mission that the robot should perform are the main tasks that should be performed by its navigation system. This last task informs which path the system of motion control should perform.

One of the most important information for a mobile robot is the distance measurement to barriers. This information can be obtained through various sensors, among them images from the environment might be obtained. In possession of this distance, decisions such as location, overpass obstacles and others can be made. One of the ways to perform distance measurement is through digital images. Image stereo systems are made with two images taken simultaneous, or not, of the same scene, similar to animals that have two eyes. With this system you can recover the depth information in the images acquired [1].

This work aims to develop a system of distance measurement using computer vision to use in modules of perception in mobile robots.

The development of a computer vision system is of great importance to complete tasks by robots, with special regard to navigation of autonomous robots in unfamiliar environments. This statement is true, since vision is the primary sense used by humans and that guides their daily lives, in their work and their leisure time. There is a tendency of current mobile robots incorporate vision systems. As the use of stereo systems allows the measurement of depth, it is going to be useful in the task of locating and overpass obstacles. It should be mentioned that several other algorithms can worked together with the vision system in stereo, using the same cameras and images acquisition system [4]. This work is organized as bibliographic review from Transformed SIFT and Depth calculation using Digital Image Processing in the second part, there is the methodology used to achieve the results in the third part and in the forth there are the results of the research.

II. Bibliographic Review

The main methods to execute the implementation and to understand this system are the transformed SIFT and the use of stereo images to calculate the coordinate z distance.

A. SIFT Method

SIFT is a robust method to extract and describe keypoints. The features extracted by this technique are invariant to translation, rotation, and scale and partially invariant to changes in lighting and perspective [2] [3] [5]. Basically, the algorithm is performed in four steps:

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_44,

- Detect extremes: it is a search on all scales and locations of images with different Gaussian filters (DOG - difference of Gaussian) to identify interest points with invariance in scale and rotation;
- Locate keypoints: the places where there is an extreme, sets a detailed model to determine the exact location and scale. Keypoints are selected based on stability measures;
- Define orientation: set up an orientation for each keypoint using the local image gradient and;
- Describe keypoints: measure the local gradient of each keypoint, based on their neighborhood.

These measurements are transformed to a representation that is tolerant to significant levels of distortion and changes in lighting. In possession of the keypoints captured from two images, they are able to be compared with Euclidean distance calculation and find its correspondent. Therefore, it is possible to trace points in images, even in real time and make the mosaics' composition.

B. *Depth calculation using Digital Image Processing*

A stereo vision system is scheme in which two images of the same scene are acquired under different conditions in order to be able to make measurements in three dimensions. Animal's vision system accomplish this task easily, that is why animals generally have two eyes. Therefore, the system used is shown in Figure 3 and Figure 4. It consists of two webcams with focal length f equidistant from one source at a distance b/2. Cameras from the left and right are called l and r, respectively [1]. When viewing an object, every point of its visible contour has a coordinate (x, y, z) in physical space in relation to the original localization between the left and right lenses. To perform the coordinates x, y and z measurement are used the following equations (1).

$$x = b\frac{(x_l + x_r)/2}{(x_l - x_r)}; y = b\frac{(y_l + y_r)/2}{(y_l - y_r)}; z = b\frac{f}{x_l - x_r} (1)$$

Therefore, beyond the distance between the cameras and the focal length f of its optical assemblies, it requires the coordinates xl and xr of a point to measure its z depth. The challenge is how to find the given point in the previewed object's contour. It needs to locate the point in the image obtained with the left camera and the same point in the image from the right camera. A classic way to make this relationship between xl and xr is using correlation. Another useful way today is the use of SIFT method, and this is more robust because of its characteristics. Invariance to translation, rotation, scale, and luminous intensity of the pixels are the characteristics [1].

III. Methodology

After bibliographic review, SIFT method's algorithms are chosen and implemented. The language chosen is C/C++, using the program C++ Builder from Borland, running on Windows XP, as well as simulation platform of MATLAB as we see on Figure 1.

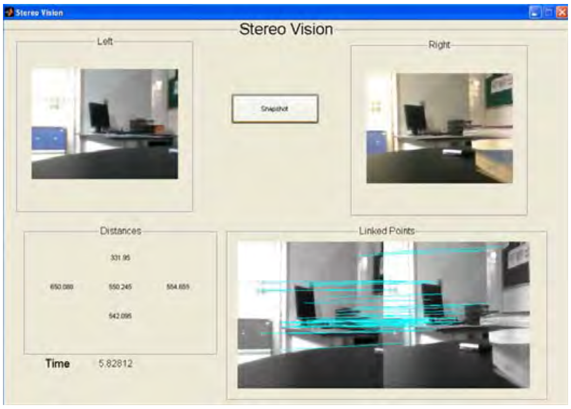


Fig. 1. User Interface.

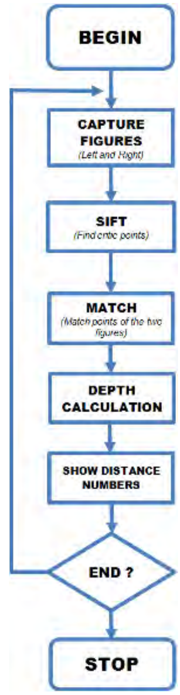


Fig. 2. Process fluxogram.

A stereo system images acquisition must be developed with two cameras and specific routines for images capturing. The depth calculation process can be seen in Figure 2.

Routines for distance measurement should be implemented from the SIFT method and the acquired images. The system calibration is made by using pattern images such as checkerboards with distances, inclinations and rotations known. Tests using standard images, controlled environments and complex should be also made. From the results obtained, the system is evaluated and the results and evaluations must be published in scientific events.

IV. Results

To perform the preliminary tests with the implemented program, two cameras are used with a distance of 120 mm from center to center. Two images are captured, Figure 3 and 4, one from the left camera and one from the right respectively. It is observed in these figures that there are some main objects: the computer monitor and the paper holder, the first is nearer to the camera than the second. When applied the SIFT method to these images and made a comparison between the keypoints found in each image, we obtain the results shown in Figure 5. The Figure 6 shows the 5 subregions. The result is the average of each keypoints' depth distance in the 5 subregions. They are: left = 650.08mm; right = 550.24mm, top = 331.95mm, bottom = 542.09mm, center = 550.24mm. The focal length f of the cameras used is unknown, but it might be calculated experimentally. The distance b between the cameras used is 120mm. It is able to be seen with the result that the subregion where the monitor is, the subregion in the center, has an average depth lower than the subregion where the paper holder is, the right subregion. This entire process occurs in 5.8 seconds and can be observed through a graphical user interface in Figure 3.

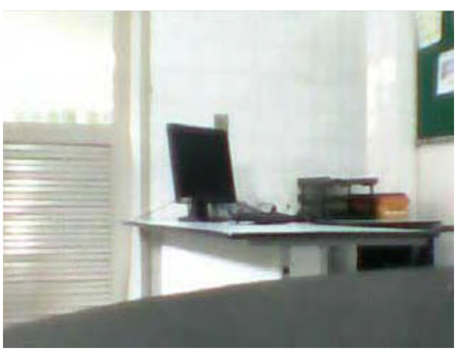


Fig. 3. Picture taken from the left camera.

It was realized a second test, two more pictures from the same scene were captured and were processed with by the system. The obtained results from the second test are: left = 535.10mm; right = 730.59mm, top = 321.28mm, bottom = 481.55mm, center = 554.77mm. We obtain the numbers from the distances and this make possible to conclude that the system it is working because we moved the cameras away from the scene and the numbers are higher than the obtained in the first test. There are some times which the program does not find keypoints in certain parts of the images. This implies a zero value in one or more subregions displayed in the interface.

Fig. 4. Picture taken from the right camera.

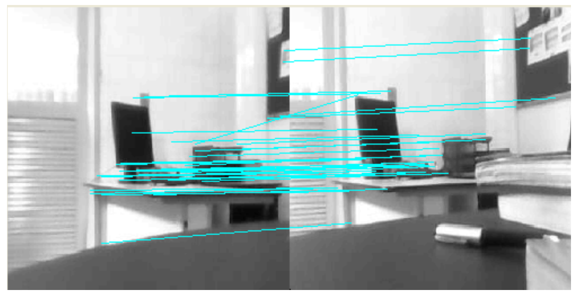


Fig. 5. Result of transformed SIFT in the first application.

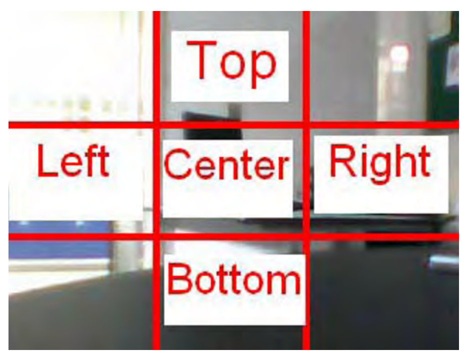


Fig. 6. Defined subregions.

V. Conclusion

This work aims to develop a system based in computer vision to obtain the depth distance. It is developed in MATLAB with two cameras of same focal length and spaced with a known distance.

Using the system we captured two pictures from the cameras and prosecuted the images and the SIFT to get the distances. The obtained results are very promising and consistent with the expected. The simulation approaches to the real situation. It occurs because the acquired distances are very close to the original. There were realized a few tests with the system. In these tests we changed the main objects in the scene and changed the distances between the cameras and the objects. These changes was shown in the interface, as we saw in Figure 3. The numbers were coherent with the reality. If we moved back the cameras, we obtained higher numbers in the interface. If we moved just one side of the cameras, the numbers changed just in the area we modified.

The next steps to be performed are to execute a quantitative test to calibrate the measurement system and embed the system in a mobile robot.

Acknowledgment

We thank to Elder Teixeira.

We thank to the laboratories LEM and ITTI from IFCE-Instituto Federal de Educação, Ciência e Tecnologia do Ceará and the laboratory LESC from UFC-Universidade Federal do Ceará.

References

[1] SIEGWART, R. and NOURBAKHSH, *I. R. Introduction to Autonomous Mobile Robots,* Bradford Book, 2004.

[2] ANGELI, A., DONCIEUX, S., MEYER, J.-A. and FILLIAT, D, *Incremental vision-based topological slam, Intelligent Robots and Systems.* IEEE/RSJ International Conference, 2008.

[3] LOWE, *D. G. Distinctive image features from scale-invariant keypoints.* Int. J. Computer Vision, 2004.

[4] TOON GOEDEMÉ, MARNIX NUTTIN, T. T. e GOOL, *L. V. Omnidirectional vision based topological navigation.* International Journal of Computer Vision, 2007.

[5] VALGREN, C., LILIENTHAL, A. e DUCKETT, *T. Incremental topological mapping using omnidirectional vision.* Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference, 2006.

Dynamic of New Technologies in Switching Power Devices IGBTs

Christian Ayala Esquivel
Jorge Ochoa Somuano
Isidro Moctezuma Cantorán
Universidad del Mar Campus Puerto Escondido (UMAR)
Carretera San Pedro Mixtepec-Sola de Vega, Cd. Universitaria, Puerto Escondido, Oaxaca,71980 México
{christiancae,ochoa,moctezuma}@zicatela.umar.mx

Abstract: **This paper presents the results from a study on the dynamic new technologies that present under various conditions when the current varies in hard switching, these data were obtained in response to use as a case study of the isolation gate bipolar transistor, which is a power semiconductor device.**

I. INTRODUCTION

The proper selection of a power electronic device, one must consider aspects such as cost, the parasitic elements and the power dissipated. This latter aspect directly influences the efficiency of the converter, especially topologies that used hard switching on the switches. Therefore, correct evaluation of the losses generated in the type of converter and its operating conditions are very important to optimize the circuit design. The information collected for the assessment of the manufacturer's data sheet, is often very limited or incomplete and represents approximate values, taking into account the tolerances of the devices commonly present. Another drawback stems from the fact that these data do not lend themselves to comparison between devices from different manufacturers, since they were obtained under different operating conditions, temperature, current level, gate resistances and so on. The comparison of various technologies is in terms of conduction losses and switching. In this paper, we analyzed only the case of hard switching, since it generates the most losses during transients. Among the power electronic devices, the IGBT of insulated gate is currently the DSEP power semiconductor device with the widest range of applications and possibilities for optimization in the future. The IGBT can be considered a hybrid between a bipolar transistor and a field effect transistor (MOSFET) obtaining the best features of both. The IGBT copied the best features of MOSFET: simple control circuit, safe working area and high peak currents; and the bipolar transistor [1].

II. MANUFACTURING TECHNOLOGIES

Currently, in the market there are three IGBTs from manufacturing technologies: Punch-Through, Non-Punch-Through and Field-Stop [9] - [11].

A. Technology "Punch-Through (PT)"

This technology IGBT has a non-homogeneous structure. In this device fabrication technology is built on a (p+) type substrate thickness (≈300µm). The (n-) layer is obtained by epitaxial growth and is relatively thin. "Fig. 1" shows a very thin (n+) layer between the base and emitter of the internal pnp named buffer, which limits the expansion of the space charge region in a state of Punch-Through blocking effect and reduces the emitter injection in conducting.

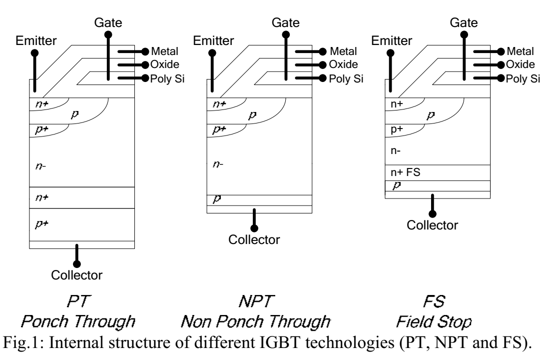


Fig.1: Internal structure of different IGBT technologies (PT, NPT and FS).

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_45, © Springer Science+Business Media B.V. 2010

The PT (IGBT) reaches low conduction losses (low resistance) through a high emitter coefficient of internal BJT, low-switching losses through a reduced lifetime [1]. This produces a carrier distribution with high slope (dp/dx) in the (n-) and a transient off current with large amplitude that decreases rapidly and depends largely on the lifetime. A device with these characteristics is the International Rectifier IRG4PH50S [11].

B. Technology "Non Punch-Through (NPT)"

The NPT (IGBT) has a structure technology originally developed by the company Siemens. This device is built in a homogeneous (n-) type substrate of approximately (220μm) wide. The issuer is developed through the implementation of a very thin layer (p+) doped and low (transparent emitter) on the back of the substrate, as shown in "Fig 1". Therefore, the IGBT is made homogeneous by the modulation of the base resistance through a low emitter coefficient in combination with a very long carrier lifetime. These characteristics lead to a homogeneous carrier distribution in the (n-) zone. Therefore, with the switch in the off position the space-charge displaces a significant portion of the charges stored on the side of the collector of internal (pnp) blocking voltages still small by checking one of the principal differences with the PT technology. Moreover, the stored charges of the emitter side of the internal (pnp) become dislodged through recombination at the transparent emitter surface. Both effects mentioned ensure low switching losses of NPT type IGBT without the use of lifetime control. The structure was electrically characterized by a transient current in the tail-shaped switch in the off position of reduced amplitude that decays slowly and is almost invariable with the temperature and does not show the effect of lifetime [1]. This is the most appropriate technology for high voltage, so most manufacturers now produce the NPT with breakdown voltages of 1200V. Furthermore, homogeneous IGBT is very robust, since it has a good performance under hard conditions and under short circuit conditions (safe operating square area), a device of this type is the IGTG11N120CN from Fairchild [10], [2] - [5].

C. Technology "Field-Stop (FS)":

The FS (IGBT) technology was developed for applications that require a saturation voltage "$V_{CE(sat)}$". This technology was recently launched by the Infineon Technologies Company (formerly Siemens Company).With this structure, conduction "$V_{CE(sat)}$" losses will by reduced, making the (n-) even thinner (≈120 μm) through the integration of an additional layer (n+) layer called Field-Stop (similar to a buffer in a PT structure) as shown in "Fig. 1". The initial problem for the manufacture of the device was handling very thin wafers. A device with these characteristics is the IKW25T120 from Infineon Technologies [9].

III. NEW GATE STRUCTURES

Lately it has been applied to all technologies, a new gate structure "trench-gate" instead of the conventional planar gate, as shown in "Fig. 2".

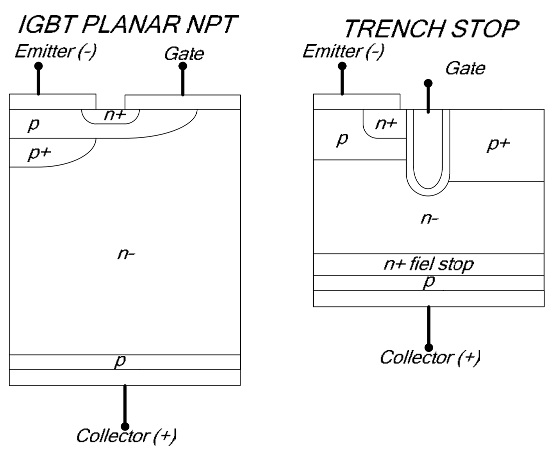


Fig. 2: Planar gate structure and trench.

This structure helps to reduce conduction losses, as the MOSFET channel is shaped vertically to the chip surface and requires a less active area. This will increase the density of cells. In short, the benefits obtained from the trench-gate structure including a decrease in the MOS channel resistance and the elimination of the JFET region [2]-[8]. The gate structure can be applied to any manufacturing technology of MOS devices. The combination of Field-Stop technology with trench gate structure led to a new Trench Stop IGBT (TS).

IV. SWITCHING MODES

The following describes the different switching modes that can be put a switch-controlled power converter in an application [1].

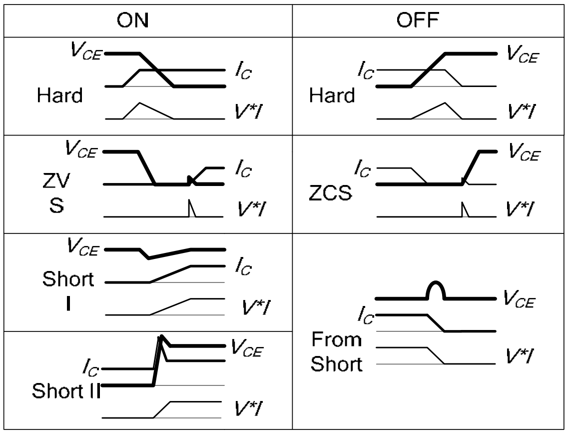


Fig 3: Different switching converter.

1) Hard Switching: The switching of a circuit breaker is defined as hard when there is simultaneous presence of current and voltage in the time period that lasts the switching phase. It can occur in the ignition and off the device and is characterized by high losses "Fig. 3".

2) Soft Switching: The switching of a circuit breaker is defined as mild when the voltage or current is zero at the start of phase switching. It can occur in the ignition and shut off and is characterized by low losses "Fig. 3". The soft switching is classified into:

Zero voltage switching or ZVS: This type of switching occurs only during power when the voltage is zero before it starts to flow a current through the device.

Zero current switching or ZCS: This type of switching occurs only during shutdown, when the current is zero before the device starts to block a voltage. [3], [4].

3). Short circuit: The short is a type of switching is not desirable in a converter, which in principle corresponds to a hard switching on the ignition under certain conditions the external circuit, being the most frequent short-circuiting the load. This switch is characterized by very high losses that lead to the device to the limits of its safe operation. Depending on the time in which it occurs, there are two types of short circuits:

Short circuit type I: the short circuit occurs during the locking phase, before the device turn on "Fig. 3".

Short circuit type II: the short circuit occurs during the driving phase, before the device off in the "Fig. 3".

V. TEST CIRCUIT

The test circuit used for the characterization in hard switching converter is a “DC/DC” type gearbox or also called chopper. The circuit diagram is shown in "Fig. 4", including the device under test “DUT” and the auxiliaries.

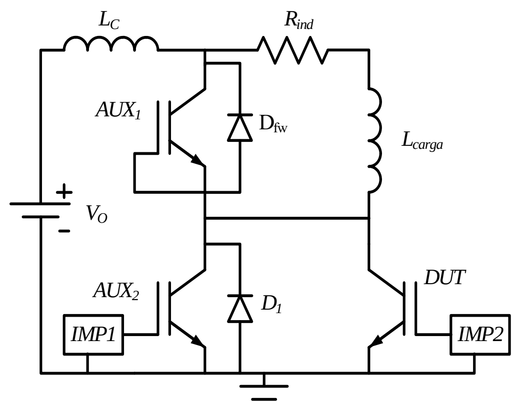


Fig 4: Test circuit for hard switching.

The “AUX_1” auxiliary device works on one side a free movement diode in the normal operation tests and on the other side as for short-circuit tests. The “AUX_2” auxiliary device is placed parallel with the device being tested and it circulates the initial current of the inductive load to the level of current necessary for testing. Thus, the “DUT” leads the minimum time required and avoids the effect of internal heat losses by conduction (conduction current by saturation voltage), obtaining a better control junction temperature of the device. “L_C” is the parasitic inductance wiring. “R_{ind}” is inductance resistance load. “L_{carga}” is load inductance where the energy is stored for the current test. “IMP1” and “IMP2” are drivers that perform the action to control the activation of auxiliary devices under test.

The nominal test circuit: collector current I_C=25A, gate resistance R_G=22Ω, voltage V_O=600V, wiring inductance L_C=116nH, gate voltage V_{GG}=15V and junction temperature T_i=30°C.

VI. EXPERIMENTAL RESULTS

In this section we present the results in three devices with hard switching characterized by varying the parameters of the I_C current. Energies are shown only in the ignition and shutdown, which were calculated by integrating the instantaneous power dissipated during the switching transient. “Fig. 5” shows tensions and currents of V_{CE}, I_C respectively as well as the instant potency calculate as a product of this and the instantaneous power is calculated as a result of these and the energy obtained by integration of the power, see “Fig. 5”.

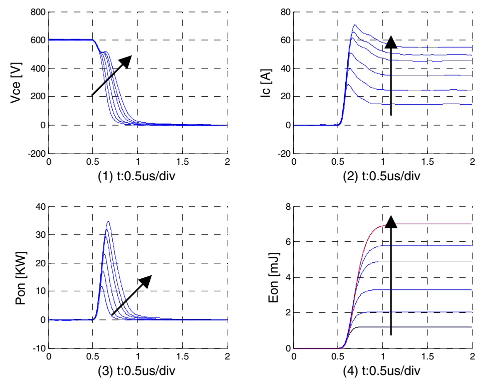


Fig. 5: Behavior of the voltage V_{CE}, current I_C, power dissipated energy P_{ON} and E_{ON} in the ignition hard switching, “IKW25T120”.

In "Fig. 5" the ignition switch is displayed, where the collector-emitter voltage V_{CE} generating step decreases due to the parasitic inductance. The inductance affects the slope of the current “di/dt”. While the current spike is caused by the diode reverse recovery is the test bench and on a current plotted peak is noticed as a result. This displays an increase in power and the energy dissipated during the on phase, on the device from Infienon Technologies.

"Fig. 6" also demonstrates the Fairchild Company power switching device, where the collector-emitter voltage V_{CE} generating the same step decreases due to the parasitic capacitance that is similar to above, with a much greater slope. The current spike is caused by the same reverse recovery diode containg in the testbank causing the same effect. This displays an increase in power and the energy dissipated as it switches to the on position. Infineon Technologies in hard switching consumed to the maximum power in testing (6mJ) as well as International Rectifier. While Fairchild only consumed (5mJ).

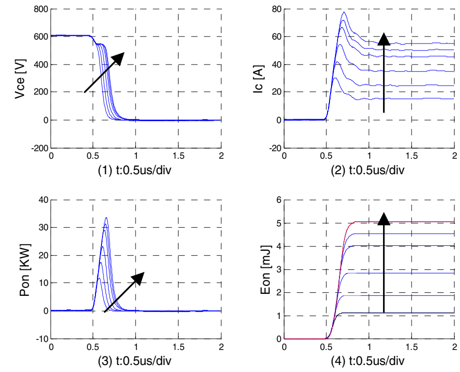


Fig. 6: Behavior of the voltage V_{CE}, current I_C, power dissipated energy P_{ON} and E_{ON} in the ignition hard switching, “IRG4PH50”.

Lastly "Fig. 7" shows the switching on the International Rectifier were, where the collector-emitter voltage V_{CE} generating the same step decreases due to parasitic inductance. This displays an increased potency similar to that of Infineon and dissipated energy at the time of power up the device from company International Rectifier. Note that the time scales are similar and can see very well the best benefits in switching the ignition.

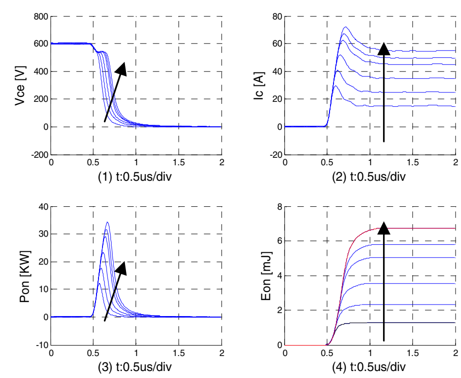


Fig. 7: Behavior of the voltage V_{CE}, current I_C, power dissipated energy P_{ON} and E_{ON} in the ignition hard switching, “IGTG11N120CN”.

The case of commutation when you want to turn off a device is different, see “Fig. 8, Fig. 9 and Fig. 10”.

"Fig. 8" again this switching device Infineon, is shown off the device showing the effect of parasitic capacitance when the device off. Subsequent voltage slope (dv/dt). This also leads to increased energy in the off causing losses at high frequency. About the maximum is (4mJ).

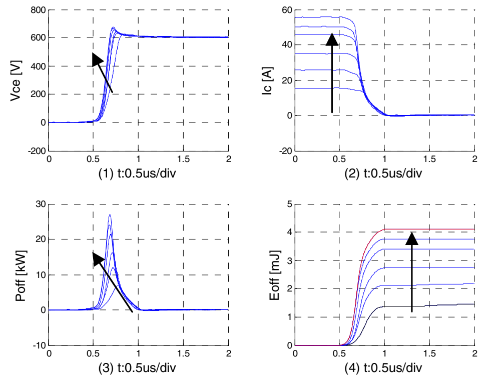


Fig. 8: Behavior of the voltage V_{CE}, current I_C, power dissipated energy P_{ON} and E_{ON} in the ignition hard switching shutdown, “IKW25T120”.

In "Fig. 9" shows the switching behavior in shutdown where the device is the most deficient, approximately consumes the energy (23mJ) of Fairchild Company, showing that this device can’t be used at high frequencies have a very slow shutdown.

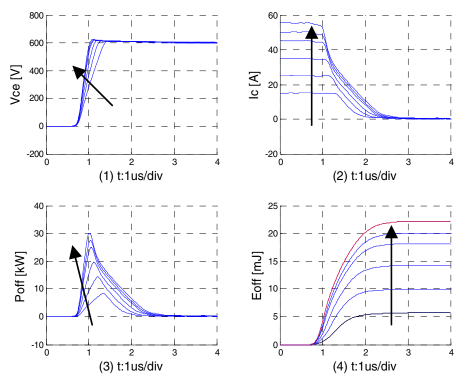


Fig. 9: Behavior of the voltage V_{CE}, current I_C, power dissipated energy P_{ON} and E_{ON} in the ignition hard switching shutdown, “IRG4PH50S”.

Unlike earlier in the "Fig. 10", is shown device off from International Rectifier, this device presents a good behavior by having a relatively small tail off and has an energy of (2.3mJ) which is about half the energy the company of Infineon Technologies.

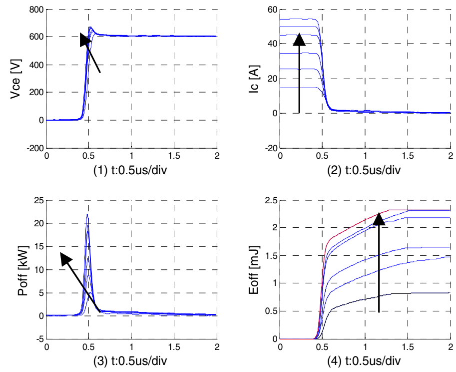


Fig. 10: Behavior of the voltage V_{CE}, current I_C, power dissipated energy P_{ON} and E_{ON} in the ignition hard switching shutdown, “IGTG11N120CN”.

VII. CONCLUSION

The experimental results on tests (hard-switching), the three technologies analyzed show that the ignition energies during on and off switching is different as the parameters vary. There are differences in the level of impact it can have a parameter variation in each of the devices, those belonging to different technologies. Considering the results, is necessary extend the experimental characterization from hard switching, changes in current, voltage, temperatures of union, gate resistance and parasitic inductance. The optimization of devices, increases efficiency and therefore energy saving.

VIII. FUTURE WORK

Obtain the complete mathematical model of hard switching behavior and maximize the efficiency of the tested power devices.

REFERENCES

[1] J.Aguayo., “Análisis Comparativo de Transistores IGBT tipo PT y NPT en Diferentes Modos de Conmutación”, Tesis de Maestría, CENIDET, Cuernavaca, México, 2000.

[2] H. Ruething, F. Umbach, O. Hellmund, P. Kanschat, G.Schmidt, “600V-IGBT3: Trench Field Stop Technology in 70μm. Ultra Thin Wafer Technology”, Memorias de la conferencia ISPSD, 2003, p.

[3] T. Laska, G. Miller, M. Pfaffenlehner, P. Türkes, D. Berger, B. Gutsmann, P. Kanschat, M. Münzer, "Short Circuit Properties of Trench-/Field-Stop-IGBTs – Design Aspects for a Superior Robustness" Memorias de la conferencia ISPSD, 2003, p. 1-4

[4] H. Iwamoto, H. Kondo,S. Mori, J. Donlon, A. Kawakami, "An Investigation of Turn-off Performance of Planar and Trench-Gate IGBTs under Soft and Hard Switching", Memorias de la conferencia IAS, 2000, p. 1-6.

[5] R. Mallwitz,R. Tschirbs,M. Pfaffenlehner, A. Mauder, C. Schaeffer, "1700V Trench IGBT Modules", Memorias de la conferencia PCIM, 2001, p. 1-6.

[6] X.Kang, A. Caiafa, E. Santi, J. Hudgins, R. Palmer, "Characterization and Modeling of High-Voltage Field-Stop IGBTs", IEEE Trans. Industry Applications, vol. IA-39/4, p. 922-928, 2003.

[7] R. Kraus, M. Redding, K. Hoffmann, "The Short-Circuit Behaviour of IGBTs Based on Different Technologies", Memorias de la conferencia, EPE, 1995, p. 1_157-1_160.

[8] S.Azzopardi,C. Jamet, J.M. Vinassa, C. Zardini, "Switching Performances Comparison of 1200V Punch-Through and Punch-Through IGBTs under Hard-Switching at High Temperature", Memorias de la conferencia IEEE PESC, 1998, p. 1201-1207.

[9] www.infineon.com

[10] www.fairchildsemi.com

[11] www.irf.com

Anisotropic Percolating Pathways in the Thin Films of Polymeric PEDT/PSS Complex and their Relation to the Electrical Conductivity as Revealed by the Mesoscale Simulation

Toomas Kaevand[a], Jaan Kalda[b], Andres Öpik[a], Ülo Lille[a1]

[a]*Department of Material Sciences, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia*

[b]*Institute of Cybernetics, Tallinn University of Technology, Akadeemia tee 21, 12618 Tallinn, Estonia*

Abstract. The MesoDyn simulation code supplemented with the percolation and cluster analysis concepts has been used to generate the subcritical percolating networks of PEDT beads of the coarse-grained PEDT/PSS complex in a range of PEDT/PSS mass ratio up to 20. Based on the comparison of the percolation parameters in three perpendicular directions, the anisotropy of the networks has been determined and its correlation with the macroscopic electrical conductivity has been evaluated.

I. INTRODUCTION

The term “percolation pathway” has been often used in the interpretation and generalization of the experimental data in the research of the transport properties of electrically conducting conjugated polymers, e.g. [1-3]. However, to our best knowledge, there are no reports devoted to the analysis of this topic (except for our efforts as shown below). At the same time, the respective studies of the carbon nanotubes are progressing [4-5].

Poly(3,4-ethylenedioxythiophene) (PEDT) blended with poly(4-styrenesulphonate (PSS) is one of the most intensively studied inherently electrically conducting conjugated polymer, and its thin films are widely used in several industrial applications [6-11]. The properties of these films depend on their morphology, which, in its turn, is a function of the PEDT/PSS mass ratio and of the processing conditions. The morphology and the factors affecting it have been studied by a wide range of physical methods, including electrical measurements, and supported by atomic scale modeling [12-28].

Current study is largely motivated by a recently found convincing evidence on the highly anisotropic nature of the electrical conductivity of the PEDT/PSS complex (1:6 by weight) [2,29 and citations therein]. The main conductivity mechanism in the thin PEDT/PSS film at 300K is the variable range hopping (VRH) of the charge carriers from one conducting center to another center [29-30].

In this report, the authors continue their earlier efforts to simulate the behavior and the transport properties of the PEDT/PSS complex using dynamic density field theory (DDF) [31-32] and principles of classical percolation theory [33]. The simulated morphologies have been given in the form of the density fields as ordered systems of the coarse-grained polymer molecules (mesomolecules represented as Gaussian chains of beads) in a grid. The percolation phenomenon has been treated on a discrete lattice and also in the continuous space, and the relevant parameters of the correlated conducting network of the PEDT particles have been calculated [34-35].

The aim of this research is 1) to establish the degree of anisotropy of the percolation pathways and its dependence on the composition of the complex; 2) to determine the relationship between the percolating current-carrying backbone density (strength) and the electrical conductivity; 3) to clarify the limits of the used simulation approach.

This is not an easy task, keeping in mind that the percolating pathways are wispy, tenuous, insubstantial fractal objects [36] that have no characteristic length scale and furthermore, little is known about anisotropic percolation [37].

In this report, we demonstrate the anisotropic nature of percolating pathways in a wide range of the composition of the complex, and discuss their relation to the macroscopic electrical conductivity.

II. METHODOLOGY

The methodology of the mesoscale simulation in a 32^3 cubic lattice using the MesoDyn simulation code (from Accelrys Inc. Material Studio Library Bundle, MS 4.4) and the treatment of the generated morphologies in the framework of the isotropic percolation concept is described in detail in our previous reports [34-35]. The topology of the PSS mesomolecules corresponds to that of random copolymers (if not stated otherwise). Due to the phase separation in the PEDT/PSS systems, the lamellar *anisotropic* structures are formed. However, we are forced to use *isotropic* percolation model, due to the lack of well-developed anisotropic models.

To facilitate the further reading, some details on the generation of the percolation pathways are given below.

Our self-developed cluster analysis algorithm is used to generate the subcritical percolating networks. The lattice site occupation probability p (not distinct from the volume fractions of the PEDT beds) is taken to be very close to the effective percolation threshold $p^*_{c,}$, at the distance $\Delta p = p^*_c - p$. The procedure is as follows. First, MesoDyn is used to generate a density field $d(X,Y,Z)$ on a 32^3 lattice (i.e. each cell with coordinates X, Y and Z is characterized with a certain density d). Further, a cite (cell) is postulated to be

[1]E-mail addres lille@chemnet.ee.

Ü. Lille thanks Estonian Science Foundation for financial support (Grant No 6633).

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_46, © Springer Science+Business Media B.V. 2010

occupied, if its density is higher than $(1 - y)d_{\max}$, and vacant otherwise; here, y is a control parameter, and $d_{\max}$ is the maximal density. The parameter y is increased slowly, until the global conductance along the surviving sites appears; at this moment, the critical density $d_c = (1 - y)d_{\mathrm{dmax}}$ is recorded. The density interval $|d_{\max} - d_c| \div d_{\max}$ will be referred to as the conducting window. The accuracy of the parameter y (4 significant digits) defines the value of the sub-criticality parameter $\Delta p = 10^{-4}$. Finally, the obtained subcritical network of the PEDT beads is used to determine the number of cells in the conducting window (N_{con}) and in the percolating pathway (N_p). Removing the "dead ends" one gets the number of cells in the charge carrying backbone N_b. The "dead" ends are the parts of the infinite cluster connected to it only by one site (they lead nowhere). Ratio of these three values to the total number of cells gives the (effective) critical probability p^*_c and the densities of the percolation path $P(p)$, and of the backbone $P(p)_b$. Similarly, one can calculate the fraction of the singly connected ("red") sites f_r. In principle, all current passes through these "red" sites (however, see below section III. C).

The percolation path is a cluster (effectively infinite) connecting the opposite sides of the 3-D lattice. Due to the system proximity to the percolation threshold, the correlation length is certainly larger than the lattice side, i.e. our systems are scale invariant (or self-similar) [37-38]. We generate the percolation path for all three directions *X* ,*Y*, *Z* independently. In what follows, the respective pathways are termed as the percolation paths for the first, second and third direction; the sub-indices of the parameters designate the used direction. Note that the first-appearing path does not necessarily always correspond to the same Cartesian coordinate axis.

III. Results

III. A. Initial Systems And Their Macroscopic Electrical Conductivity

For the studied systems, the PSS to PEDT mass ratio (termed as S/T) ranges from 1.4 to 20; the corresponding PEDT volume fraction θ_T varies from 0.495 to 0.05. In this set of experiments, the doping level is kept constant (0.25, i.e. one PSS anion per four PEDT cycles if not stated otherwise). The macroscopic electrical conductivity (σ) of the studied systems ranges from ca 1 to 6×10^{-5} S cm^{-1} (Tab. 1) [3,19]. These data are correlated with the θ_T values by the linear equation expressing the exponential function $\sigma_i = A \exp(b\theta_T)$:

$\ln \sigma = 19.27\ \theta_T - 8.651$; (items 1-7) $R^2 = 0.86$ (1);

pay attention that $b>0$. Naturally, ln σ versus S/T ratio correlates with $b < 0$. Such an empirical law reflects the fact that the macroscopic conductivity values range in a very wide (exponentially wide) range of values. It should be also stressed that the morphologically induced conductivity variations for systems with a fixed S/T value are ignored by this law.

TABLE I

PSS/PEDT mass ratios (S/T), volume fractions of the PEDT beads (θ_T) and macroscopic electrical conductivity (σ) S cm^{-1}.

1)from the data [3,19] using correlation function $\ln \sigma = 4.1738 \ln\theta + 2.9248$, $R^2 = 0.961$.2) PSS is modeled as an alternating copolymer; 3) as in 2, but unassociated PSS chain is ca 3-fold longer; 4) under shear 0.001 ns^{-1}, repulsive interaction 2.6 RT; 5) doping level 0.03.

Item	S/T	θ_T	10^{-3} x σ
1	1.4	0.495	600
2	2.5	0.333	400
3	4.2	0.193	20 1)
4	6	0.143	5
5	8	0.110	2 1)
6	10	0.091	1 1)
7	20	0.05	0.06
8 2)	2.5	0.33	400
9 3)	2.5	0.33	400
10 4)	2.5	0.33	400
11 5)	2.5	0.33	?

Item 11 is introduced to give an idea of how the change of the doping level affects the results. This item is not included into the correlation analysis shown below. In general, its behavior is rather similar to that of others.

Note that in all experiments, the macrophase separation of the neutral host mesomolecule (unassociated PSS) is observed (as evidenced by the order parameter values over 0.1, except for the item 7). The respective values of the guest PEDT beads remain in the range of 0.005-0.01, characterizing the microphase separation. The behavior of the polyanionic PSS beads remains between these two extremes.

III. B. Charge Carriers Hopping And The Distance Between The Conducting Sites

The hopping mechanism implies that the charge carriers tunnel between the conducting sites. Neglecting the fluctuations of the barrier height, the tunneling probability (and the local conductivity) decreases exponentially with the distance between the conducting sites**.** Therefore, there are very strong fluctuations of the local conductivity, and the global (macroscopic) conductivity along some axes (e.g. *X*) can be estimated as the "bottleneck" conductivity σ_c [33], i.e. such a largest conductivity σ_c that the region $\sigma(x,y,z) > \sigma_c$ percolates in the direction of the given axes (*X*).

As argued above, the local conductivity is proportional to the exponent of the distance between the conducting sites, which in its turn is proportional to $d^{-1/3}$, and at the bottleneck site, to $d_c^{-1/3}$.

This "bottleneck model" reflects the physical nature of the problem with a reasonable accuracy. However, it is still an approximation, for which the resistance of the sites, other than the bottleneck site, is completely neglected. In reality, however, this is not correct. Another drawback of the bottleneck model is that for small samples (like our 32^3), the statistical fluctuations of the bottleneck value are rather large. Therefore, it can be argued that the *average* hopping length, as reflected by the average field density within the conduction window $\langle \Delta d \rangle$, can serve in our case as an alternative measure with a lesser statistical variance. To summarize, d_c is the lower bound of the conducting window, and $\langle \Delta d \rangle$ is the median value of it.

Hence, it is expected that the global conductivity is proportional to $\exp(bd_c^{-1/3})$, or to $\exp(b\langle \Delta d \rangle^{-1/3})$, respectively, where b is an unknown constant; recall that d_c and Δd are the bottleneck field density and the average density of the conducting window of the PEDT-beads, respectively.

TABLE II
Maximal (d_{max}) and critical (d_c) field densities in the perpendicular directions 1, 2, 3.

Item	d_{max}	d_c 1	d_c 2	d_c 3
1	0.9542	0.809	0.7960	0.791
2	0.8452	0.638	0.6012	0.601
3	0.6845	0.407	0.3971	0.397
4	0.5347	0.397	0.1930	0.0812
5	0.4301	0.0794	0.0778	0.0763
6	0.3536	0.0694	0.0694	0.0691
7	0.1191	0.0495	0.0495	0.0494
8	0.573	0.467	0.466	0.443
9	0.555	0.456	0.454	0.433
10	0.541	0.447	0.421	0.421

The simulation data (see Tab. II), indeed, confirm the expected correlation, for all the three directions (items 1-10):

$\ln \sigma = -4.8767\, d_{c1}^{-1/3} + 4.3383$, $R^2 = 0.8215$ (3),

$\ln \sigma = -5.1979\, d_{c2}^{-1/3} + 5.0938$, $R^2 = 0.90$ (4)

$\ln \sigma = -4.138\, d_{c3}^{-1/3} + 4.0129$, $R^2 = 0.65$ (5)

Note that excluding the special cases 8,9, and 10 improves the correlation for the 3rd direction, resulting in $\sigma \propto \exp(-4.5045\, d_{c3}^{-1/3})$ and $R^2 = 0.89$.

Pay attention that $b < 0$, i.e. $\ln \sigma$ decreases when $d_c^{-1/3}$ increases (which corresponds to decreasing d_c).

Similarly good (even slightly better) correlation is observed when the average conducting window density $\langle \Delta d \rangle$ is used instead of d_c (items 1-7):

$\ln \sigma = -7.107 \langle \Delta d \rangle_{c1}^{-1/3} + 5.5708$, $R^2 = 0.87$ (6)

$\ln \sigma = -7.358 \langle \Delta d \rangle_{c2}^{-1/3} + 6.071$, $R^2 = 0.90$ (7)

$\ln \sigma = -7.4394 \langle \Delta d \rangle_{c3}^{-1/3} + 6.2799$, $R^2 = 0.91$ (8)

Qualitatively analogous relationship (however, just for a single direction) was observed in our recent report [39].

Pay attention to the fact that these correlation analysis results are relatively insensitive with respect to the change of the fitting parameter (d_c vs $\langle \Delta d \rangle$). Further, the values of the exponential factors for different directions are rather close and therefore, the expected high anisotropy, i.e. high ratio of the conductivity in parallel to the substrate surface plane ($\parallel$, in-plane) and perpendicular to it ($\perp$, out-of-plane) cannot be identified. An exceptional case is the item 4 (S/T = 6), for which this ratio achieves a value of about 25. Thus, these field density values (d_c and $\langle \Delta d \rangle$) are unable to reflect the electrical anisotropy in the set of studied systems. In what follows we argue that this fact can be explained by the imperfections of the bottleneck model and that the strong anisotropy of the global conductivity can be ascribed to the strong anisotropy of the topological structure of the percolation clusters at criticality (which is, indeed, observed in our simulations).

III. C. ANISOTROPY OF THE PERCOLATION PARAMETERS AND THE MACROSCOPIC ELECTRICAL CONDUCTIVITY.

Below are compared the percolation parameters characterizing the subcritical networks generated for the samples with different S/T ratios (see Tab III).

The relatively high p^*_c values give an evidence that a lot of cells are located in the finite clusters. A rather modest anisotropy is observed.

One can see that there is a considerable anisotropy for $P(p)$, as evidenced by the fact that the values for the first two directions are noticeably lower than in the case of the third direction. This means that more sites are needed to circumvent the insulating PSS beads, i.e. in order to percolate in the third direction.

Remarkable is that $P(p)$ values are rather close to the $P(p)_b$ values. Hence, the subcritical network has very few dead ends. Note also that according to our previous report [34], $P(p)$ behaves according to a power law, $P(p) \propto |p-p_c|^{\beta}$, where the critical exponent for the infinite cluster density $\beta > 1$. We also showed there that our systems correspond to the *correlated* percolation, and the power law of $P(p)$ is observed up to $p = 1$. This is important, because in the real systems $p - p_c$. can take non-small values.

The most anisotropic quantity is $P(p)_b$: over the full range of the mass ratios and simulation conditions one can observe the dependence of $P(p)_b$ on the direction. The average value of the charge-carrying backbone strength in the third direction exceeds that of in the first direction ca four-fold.

Being guided by this finding, we also studied the correlation between $P(p)_{b1}$ and the macroscopic conductivity; a moderately correlated linear relationship was found (items 1-10):

$\ln \sigma = -84.79\, P(p)_{b1} - 0.707$, $R^2 = 0.51$ (9).

This correlation improves somewhat, if the discordant item 5 is discarded, resulting in

$\ln \sigma = -132.4\, P(p)_{b1} + 0.3641$ $R^2 = 0.62.$ (10)

Pay attention that $b < 0$, thus, the conductivity decreases very rapidly with the increase of $P(p)_{b1}$ values (e.g. 3-6 orders of magnitude at 4-fold increase of the backbone density).

Although Eqs. (9) and (10) reflect just a modest correlation, let us have a look on the consequences of them. Particularly remarkable in Tab. III is the item 4 (S/T = 6, as was used in [29]), where the ratio $\langle P(p)b_{1,2}\rangle/P(p)b_3 = 7.5$ corresponds to $\langle \sigma_{1,2}\rangle/\sigma_3 = 10^{18}$, i.e. the out-of-plane conductivity is vanishingly small. Items 2 and 9 also contrast with the remaining items: the backbone strength in the directions 1 and 2 is very similar and differ strongly from that of the third direction, leading to the $\langle \sigma_{1,2}\rangle/\sigma_3$ values of 10^2 and 10^7, respectively. From the point of view of the experimental data, these are quite reasonable results. A special case is the item 10: a shear applied in the *X* direction, resulting in the equality of backbone densities in 2nd and 3rd directions, clearly differing from that of the first one, similarly to the item 9 (in the latter, the unassociated PSS chain is ca 3-fold longer as compared to the conventional one). Note that shear (the processing tool to introduce an anisotropy) and subtle changes in topology increase the anisotropy.

TABLE III

Percolation functions in perpendicular directions (1, 2, 3) at various PSS/PEDT mass ratios (S/T) at the doping level 0.25 (if not shown otherwise) and temperature 298K.

1) PSS is modeled as an alternating copolymer; 2) as in 1), but unassociated PSS chain is ca 3-fold longer; 3)under shear 0.001 ns^{-1}, repulsive interaction 2.6 RT,; 4) doping level 0.03; 5) average from 3 parallel experiments.

Item, (S/T)	10^{-2} x p^*_c			10^{-2} x P(p)			10^{-2} x $P(p)_b$		
	1	2	3	1	2	3	1	2	3
1,(1.4)	6.26	8.24	9.19	2.47	3.69	5.17	2.23	3.45	4.87
2, (2.5)	10.2	12.5	14.8	2.25	2.95	13.4	2.15	2.76	12.8
3, (4.2)	12.6	14.0	14.0	2.76	10.5	10.5	2.56	9.95	9.98
4, (6.0)	6.67	23.8	40.4	1.63	8.12	37.9	1.44	10.4	37.7[5)]
5, (8.0)	35.5	37.4	39.5	9.27	10.7	20.7	9.24	10.7	20.3
6, (10)	36.8	35.8	37.7	6.15	6.53	15.9	5.97	6.14	15.4
7, (20)	25.5	25.5	27.0	7.36	7.36	11.6	6.10	6.13	9.95
8[1)],(2.5)	6.97	7.14	16.6	3.69	4.72	12.3	3.05	4.02	11.4
9[2)],(2.5)	6.10	6.76	12.6	0.82	1.71	10.5	0.65	1.44	9.86
10[3)],(2.5)	7.86	19.0	19.0	1.22	10.2	10.2	1.0	9.42	9.48
11[4)],(2.5)	12.9	15.2	16.7	5.1	6.0	10.5	5.02	5.92	10.4
Aver.	15.4	19.0	23.1	3.76	6.65	14.8	3.44	6.45	14.2
Stdev.	10.3	9.29	10.5	2.30	2.73	6.01	2.63	3.31	8.74

The previous paragraph 1) demonstrates an approximate character of the correlations ln σ vs $P(p)_b$ [given by Eqs. (9),(10)], and 2) clearly reflect the conductivity anisotropy.

Let us now study the structure of the charge carrying backbone (some of them are shown in Fig.). The $P(p)_b$ values are ca one order of magnitude higher than those of f_r. (see Tab. III and IV). Hence, the backbone contains a number of multiconnected pieces which provide several routes in parallel (loops). This is because the red sites connect the subsets of the finite clusters into the infinite one. Note that the ratio of the masses of the percolation path to the overall mass of all the cluster pieces is ca ¼ for the direction 1 and 2/3 for the direction 3 (compare the average values of $P(p)$ and p^*_c in Tab. III). These structural features may result in reducing the correlation strength ln σ vs $P(p)_b$.

Note that the average value of the fractions of the red bonds in directions 1 and 2 is lower in comparison to that of in the third direction, i.e. the behavior is similar to the backbone densities (Tab. IV). Hence, not all red sites act as connectors of finite clusters into the infinite one, i.e. part of them do not take part directly in the formation of the subcritical network. Therefore the role of the red sites in ambiguous.

TABLE IV

Fractions of the singly connected sites (f_r) in the perpendicular directons

Item	10^{-3} x f_r		
	1	2	3
1	5.19	7.51	10.3
2	3.75	3.84	18.1
3	4.61	15.6	15.4
4	2.01	4.70	7.48
5	1.95	1.37	7.96
6	4.58	6.32	8.88
7	17.8	17.7	24.4
8	5.19	7.51	10.3
9	2.78	4.94	17.9
10	4.00	14.8	14.8
Average	5.19	8.43	19.4

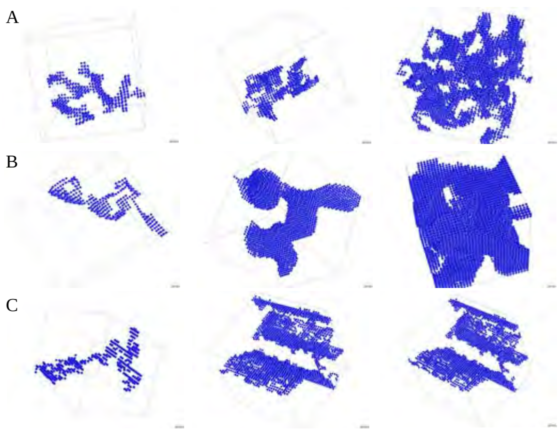


Figure. Charge carrying backbones in 3-D lattice in the directions 1,2,3 (from the left to the right).
A – item 2 (see Tab 3); B – item 4; C – item 10.

The paragraphs above are based on the random resistor model [40] (an equivalent of the hopping conductivity model [41-42]). However, in subcritical neworks used in this report the variations of the conductivity by orders of magnitude cannot be explained within such a model alone. The most accurate model would be to ascribe to each cell the resistivity based on the hopping model and to solve the system of Kirhoff's laws. This will be the subject of our further studies. Meanwhile, the reality can be approached reasonably well by combining the bottleneck model with the random resistor model: the large variations induced by the S/T values are mostly explained by the bottleneck conductivity and the random resistor model leads to a prefactor in the form of $P(p)_b$, explaining the anisotropy. Such an interpretation explains also the modestness of the correlations given in the equations (9) and (10).Therefore, the results presented above [the presence of a good correlation of the global conductivity versus $\exp(d_c^{-1/3})$, together with the strong anisotropy of the percolation backbones, as evidenced by $P(p)_b$] suggest that the overall dependence takes the form of

$$\sigma = AP(p)_b \exp(bd_c^{-1/3}), \qquad (11);$$

hence $\ln \sigma - \ln P(p)_{b1} = bd_{c1}^{-1/3} + \ln A$ (12),

and a good correlation is observed (items 1-7):

$$\ln \sigma - \ln P(p)_{b1} = -5.220\, bd_{c1}^{-1/3} + 7.8522,\ R^2 = 0.91 \quad (13).$$

This is a logical result because (12) normalizes the macroscopic conductivity to the density of the charge-carrying backbone in this way smoothing the fluctuations of the calculated data.

In all experiments, the percolation pathway densities for the first and second directions are lower than in the case of the third one. Naturally, this morphological inhomogeneity causes the same feature in the electrical conductivity. Therefore, the first two directions are related to the ‖ conductivity and the third one to the ⊥ conductivity. Note that the ‖ conductivity constitutes the main part of the macroscopic (global) conductivity.

In summary, we have treated in the framework of the isotropic percolation concept the density fields of the PEDT beads evolved as driven by the chemical potential and Langevin diffusion in the coarse-grained PEDT/PSS system. The obtained results have given evidence on the presence of the anisotropic percolation pathways in the wide range of the composition in the thin film of the complex. It is quite remarkable that such a simulation reflects the anisotropy which in fact originates from the centrosymmetric forces in the spin coating process used for the preparation of the thin film [29]. Furthermore, we have used the classical isotropic percolation concept and multiple occupancies characteristic for the correlated systems have been neglected [43]. Therefore the DFT supplemented with the classical percolation concept is a valuable tool for the studies of the transport properties of the conjugated polymers electrostatically bound to a polyanion. As the further tasks one can consider a more detailed study of the model (11) (including the analysis using the Kirhoff's laws for the network of hopping-model resistances, see above) and the research into the morphological changes induced by the annealing at different temperatures taking into account the temperature dependence of the interaction and diffusion parameters. From the other side the proper treatment of the systems on different doping levels this methodology should be complemented with quantum-mechanical calculations.

The very recent application of the most effective physical methods, e.g. magnetoresistance measurements at low temperatures had shed light to intimate details of the electrical conduction process in conjugated polymers, e.g. distinguishing the role of the isolating PSS layer and intrinsic disorder in the PEDT particles. Such details are out of grasp of the mesoscale simulation consisting itself other models and approximation (coarse- graining and Rouse dynamics of polymer chain, the use of the mean field etc).

Conclusion

Anisotropic percolation pathways have been shown to be present in the thin films of intrinsically electrically conducting conjugated polymer PEDT/PSS complex. For a wide range of the PEDT/PSS mass ratio, evidence has been given that the electrical conductivity correlates with the characterizing measures of these pathways. A model which explains the large variations of the macroscopic conductivity induced by the PEDT/ PSS mass ratio and its directional dependence on the substrate surface has been proposed. The applicability limits of the DFT theory, complemented with the percolation concepts, have been evaluated.

REFERENCSES

[1] M. Kemerink; S. Timpanaro; M. M. deKok; E. A. Meulemkamp; M. J. Touwslager, "Three-dimensional inhomogeneities in PEDOT: PSS films," *J.Phys.Chem.B* vol. 108, pp. 18820-18825, 2004.

[2] A. M. Nardes; M. Kemerink; R. A. J. Janssen, "Anisotropic hopping conduction in spin-coated PEDOT : PSS thin films," *Physical Review B* vol. 76, pp. 1-7, 2007.

[3] D. M. DeLongchamp; B. D. Vogt; C. M. Brooks; K. Kano; J. Obrzut; C. A. Richter; O. A. Kirillov; E. K. Lin, "Influence of a water rinse on the structure and properties of poly(3,4-ethylene dioxythiophene):poly(styrene solfonate) films," *Langmuir* vol. 21, pp. 11480-11483, 2005.

[4] B. E. e. a. Kilbride, "Experimental observation of the scaling laws for alternating current and direct current conductivity in polymer-carbon nanotube composite thin flms.," *Journal of Applied Physics* vol. 92, pp. 4024-4030, 2002.

[5] S. S. Rahatekar; M. S. P. Shaffer; J. A. Elliott, "Electrical. percolation in carbon nanotube dispersions: A mesoscale modeling and experimental study," *Smart Nanotextiles* vol. 920, pp. 63-69
151, 2006.

[6] S. Ghosh; O. Inganas, "Self-assembly of a conducting polymer nanostructure by physical crosslinking: applications to conducting blends and modified electrodes," *Synthetic Metals* vol. 101, pp. 413-416, 1999/5.

[7] S. Kirchmeyer; K. Reuter, "Scientific importance, properties and growing applications of poly(3,4-ethylenedioxythiophene)," *Journal of Materials Chemistry* vol. 15, pp. 2077-2088, 2005.

[8] L. B. Groenendaal; F. Jonas; D. Freitag; H. Pielartzik; J. R. Reynolds, "Poly(3,4-ethylenedioxythiophene) and its derivatives: past, present and future " *Adv.Mater.* vol. 12, pp. 481-494, 2000.

[9] X. Crispin; e. al, "Conductivity, morphology, interfacial chemistry, and stability of poly(3,4-ethylene dioxythiophene)-poly(styrene sulfonate): A photoelectron spectroscopy study," *J.Polym.Sci.B* vol. 41, pp. 2561-2583, 2003.

[10] A. B. Kaiser, "Systematic conductivity behavior in conducting polymers: effects of heterogeneous disorder," *Adv.Mater.* vol. 13, pp. 927-941, 2001.

[11] F. Louwet; L. Groenendaal; J. Dhaen; J. Manca; J. Van Luppen; E. Verdonck; L. Leenders, "PEDOT/PSS: synthesis, characterization, properties and applications," *Synthetic Metals* vol. 135-136, pp. 115-117, 2003.

[12] K. Z. Xing; M. Fahlman; X. W. Chen; O. Inganas; W. R. Salaneck, "The electronic structure of poly(3,4-ethylene-dioxythiophene): studied by XPS and UPS," *Synthetic Metals* vol. 89, pp. 161-165, 1997.

[13] K. E. Aasmundtveit; E. J. Samuelsen; L. A. A. Pettersson; O. Inganas; T. Johansson; R. Feidenhans'l, "Structure of thin films of poly(3,4-ethylenedioxythiophene)," *Synthetic Metals* vol. 101, pp. 561-564, 1999.

[14] G.Greczynski; Th.Kugler; M.Keil; W.Osikowicz; M.Fahlman; W.R.Salaneck, "Photoelectron spectroscopy of thin films of PEDOT-PSS conjugated polymer blend: a mini-review and some new results," *J.of Electron Spectroscopy and Related Phenomena* vol. 121, pp. 1-17, 2001.

[15] L. A. A. Pettersson; S. Ghosh; O. Inganas, "Optical anisotropy in thin films of poly(3,4-ethylenedioxythiophene)-poly(4-styrenesulfonate)," *Organic Electronics* vol. 3, pp. 143-148, 2002.

[16] J. Y. Kim; J. H. Jung; D. E. Lee; J. Joo, "Enhancement of electrical conductivity of poly(3,4-ethylenedioxythiophene)/poly(4-styrenesulfonate) by a change of solvents," *Synthetic Metals* vol. 126, pp. 311-316, 2002.

[17] S. K. M. Jonsson; J. Birgerson; X. Crispin; G. Greczynski; W. Osikowicz; A. W. Denier van der Gon; W. R. Salaneck; M. Fahlman, "The effects of solvents on the morphology and sheet resistance in poly(3,4-ethylenedioxythiophene)-polystyrenesulfonic acid (PEDOT-PSS) films," *Synthetic Metals* vol. 139, pp. 1-10, 2003.

[18] A. Zykwinska; W. Domagala; A. Czardybon; B. Pilawa; M. Lapkowski, "In situ EPR spectroelectrochemical studies of paramagnetic centres in poly(3,4-ethylenedioxythiophene) (PEDOT) and poly (3,4-butylenedioxythiophene) (PBuDOT) films," *Chemical Physics* vol. 292, pp. 31-45, 2003.

[19] R. R. Smith; A. P. Smith; J. T. Stricker; B. E. Taylor; M. F. Durstock, "Layer-by-layer assembly of poly(ethylenedioxythiophene): poly(styrenesulfonate)," *Macromolecules* vol. 39, pp. 6071-6074, 2006.

[20] S. Timpanaro; M. Kemerink; F. J. Touwslager; M. M. De Kok; S. Schrader, "Morphology and conductivity of PEDOT/PSS films studied by scanning-tunneling microscopy," *Chemical Physics Letters* vol. 394, pp. 339-343, 2004.

[21] J. Ouyang; Q. Xu; C. Chu; Y. Yang; G. Li; J. Shinar, "On the mechanism of conductivity enhancement in poly(3,4-ethylenedioxythiophene):poly(styrene sulfonate) film through solvent treament," *Polymer* vol. 45, pp. 8443-8450, 2004.

[22] M. M. De Kok e. al, "Modification of PEDOT:PSS as Hole Injection Layer in Polymer LEDs," in *Physics of Organic Semiconductors*, W.Brütting, Ed.; WILEY-VCH: 2005, p 451-473.

[23] C. Ionescu-Zanetti, Mechler, A., Carter, S.A., Lal,R., "Semiconductive Polymer Blends: Correlating Structure with Transport Properties at the Nanoscale," *Adv. Mater.* vol. 16, pp. 385-389, 2004.

[24] P. Vacca; M. Petrosino; R. Miscioscia; G. Nenna; C. Minarini; D. Della Sala; A. Rubino, "Poly(3,4-ethylenedioxythiophene): poly(4-styrenesulfonate) ratio: Structural, physical and hole injection properties in organic light emitting diodes," *Thin Solid Films* vol. 516, pp. 4232-4237, 2008.

[25] A. M. Nardes; R. A. J. Janssen; M. Kemerink, "A morphological model for the solvent-enhanced conductivity of PEDOT : PSS thin films," *Advanced Functional Materials* vol. 18, pp. 865-871, 2008.

[26] M. Deepa; S. Bhandari; R. Kant, "A comparison of charge transport behavior in functionalized and non-functionalized poly 3,4-(ethylenedioxythiophene) films," *Electrochimica Acta* vol. 54, pp. 1292-1303, 2009.

[27] H. Yan; S. Arima; Y. Mori; T. Kagata; H. Sato; H. Okuzaki, "Poly(3,4-ethylenedioxythiophene)/poly(4-styrenesulfonate): Correlation between colloidal particles and thin films," *Thin Solid Films* vol. 517, pp. 3299-3303, 2009.

[28] A. Dkhissi; D. Beljonne; R. Lazzaroni, "Atomic scale modeling of interfacial structure of PEDOT/PSS," *Synthetic Metals* vol. 159, pp. 546-549, 2009.

[29] A. M. Nardes; M. Kemerink; R. A. J. Janssen; J. A. M. Bastiaansen; N. M. M. Kiggen; B. M. W. Langeveld; A. J. J. M. van Breemen; M. M. de Kok, "Microscopic understanding of the anisotropic conductivity of PEDOT : PSS thin films," *Advanced Materials* vol. 19, pp. 1196-1200, 2007.

[30] C. S. S. Sangeeth; M. Jaiswal; R. Menon, "Correlation of morphology and charge transport in poly(3,4-ethylenedioxythiophene)-polystyrenesulfonic acid (PEDOT-PSS) films," *Journal of Physics-Condensed Matter* vol. 21, 2009.

[31] J. G. E. M. Fraaije; B. A. C. van Vlimmeren; N. M. Maurits; M. Postma; O. A. Evers ; C. Hoffmann ; P. Altevogt; G. Goldbeck-Wood, "The dynamic mean-field density functional method and its application to the mesoscopic dynamics of quenched block copolymer melts," *Journal of Chemical Physics* vol 106, pp. 4260 - 4269, 1997.

[32] G. H. Fredrickson In *International series of monographs on physics*; Clarendon Press ;Oxford University Press: Oxford, New York, 2006, pp. 19-22, 196-199.

[33] D. Stauffer; A. Aharony, "Introduction to Percolation Theory," Taylor&Francis: 1994; Second Edition., pp. 89-114.

[34] T. Kaevand; J. Kalda; A. Öpik; Ü. Lille; Springer: 2009; Vol. Advances in Computer and Information Science and Engineering.

[35] T. Kaevand; J. Kalda; A. Öpik; Ü. Lille, "Correlated percolating networks in the thin film of polymeric PEDT/PSS complex as revealed by the mesoscale simulation," *Macromolecules* vol. 42, pp. 1407-1409, 2009.

[36] R. Zallen; John Wiley Sons, Inc.: New York, Chicester, Weinheim, Brisbane, Singapore, Toronto, 2004, pp. 135-251.

[37] D. ben-Avraham, *Diffusion and Reactions in Fractals and Disordered Systems*, vol. Cambridge: Cambridge University Press, 2004, pp. 19-31..

[38] V. G. Gantmakher; Electrons and Disorder in Solids Clarendon Press Oxford, 2005, p 206-211.

[39] T. Kaevand; A. Öpik; Ü. Lille; T., S., Ed.; Springer: 2008; Vol. Advances in Computer and Information Science and Engineering, p 540-546.

[40] S. Kirkpatrick, "Percolation and conduction," *Reviews of Modern Physics* vol. 45, pp. 574-588, 1973.

[41] V. Ambegaokar; P. I. Halperin; J. S. Langer, "Hopping conductivity in disordered systems," *Physical Rev.B* vol. 4, pp. 2612-2620, 1971.

[42] B. I. Shklovskii; A. L. Efros, *Electronic properties of doped semiconductors* vol. Berlin Heidelberg New York Tokyo: Springer-Verlag, 1984, pp. 95-136.

[43] R. K. P. Zia; Y. Wu; B. Schmittmann, "Percolation of a collection of finite random walks: a model for gas permeation through thin polymeric membranes," *Journal of Mathematical Chemistry* vol. 45, pp. 58-64, 2009.

Predicting Trading Signals of Sri Lankan Stock Market Using Genetic Algorithms and Neural Networks

M.M.K Dassanayake
Dr. Chandima Tilakarathne
Department of Statistics, University of Colombo,Sri Lanka
maduranga@live.com

***Abstract*-This study predict the trading signals of Sri Lankan stock market using two sophisticated machine learning techniques called Genetic Algorithm (GA) and Neural Networks. These two techniques in combination predict the direction (going up or not) of the close price of tomorrow's (day t+1) 'All Share Price Index' (ASPI) of the Colombo Stock Exchange (CSE). The study period considered was from 1st November 2002 to 31st December 2008. The influential factors considered in this study represent the intermarket influence, political and environmental factors, economic stability and microeconomic factors: such as interest rate and exchange rate. A software called 'genetic model' was developed to find the optimum input variable combination that will affect the direction of tomorrow's ASPI value. Two identical neural network models called A and B were created for two different time periods, to predict the direction of ASPI of day (t+1).**

I. INTRODUCTION

Trading on the stock markets has long been considered as an alternative investment which can make fast profits. Stock markets are very dynamic, where the functionalities and the value of transactions highly depend on the other local and global factors as well as financial and non-financial factors. Hence, to obtain a profit through the stock market, it is necessary to make intelligent trading decisions. In an attempt to gain advancement in decision making, there have been innovations in stock market trading technologies and tools.

The aim of the present study is to employ such technologies (Genetic algorithm and neural network in combination) to predict the direction of Sri Lankan stock market. Initially, the genetic algorithm will be employed to find the best set of input variable combination that will affect the direction of Sri Lankan stock market. Then neural networks will be used to check the prediction ability of the best input variable combination.

The Colombo Stock Exchange (CSE) is the main stock exchange in Sri Lanka. It is one of the most modern exchanges in South Asia, providing a fully automated trading platform. As of 31 May 2008, 234 companies are listed on the CSE, representing twenty business sectors with a market capitalization of 850 billion rupees (over US$7.2 billion), which corresponds to approximately 24% of the Gross Domestic Product of the country. The All Share Price Index is the principal stock index of the CSE in Sri Lanka. ASPI measures the movement of share prices of all listed companies. Hence, this study predicts the direction of ASPI on behalf of the Sri Lankan stock market.

However, published researches related to the predictions of the CSE are very limited. Among the published work, the aim was to predict the value (price level), but not the direction. Past studies [1] showed the importance of directional prediction as the predictability and the profitability of such predictions are higher. Also there are no past studies related to predicting ASPI using these techniques (Genetic algorithm and neural network in combination) and this study may be the first of that kind. Section II of this paper describes the methodology and the techniques applied. Section III includes data description and data preprocessing. Section IV discuses how optimum input variable combination will be selected using genetic algorithm. Section V discusses how the directional prediction can be performed using neural networks. At last, section VI presents the conclusion, the limitations of the study and the improvements suggested.

II. METHODOLOGY AND THE TECHNIQUES APPLIED

There are several variables that may affect the close price of tomorrow's ASPI [2]. From them several potential input variables were selected based on inter-market influence, macroeconomic factors and news.

A preliminary analysis was carried out in order to identify the association between these potential input variables and the response variable. Graphical description, line charts and other statistical tests were incorporated, in order to identify their trends. Then the variables with trends were preprocessed again in order to clarify the associations. The variables that have high impact on the tomorrow's ASPI were selected as the 'final input variables'. Past data of these variables (six years) was transformed in to binary status using relative return function (1).Genetic Model was used to select the optimum input variable combination out of theses final input variables that will affect the direction of tomorrow's ASPI. 'Gens' of the chromosome [3] should be final input variables and the

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_47,

'individuals' should be binary past data. A fitness function was used to assign fitness values to these individuals. Genetic model should undergo several generations until the stopping criteria is met and the generation at the stopping criteria has the highest fittest individual and that also will be the combination of optimum input variables that will affect direction of tomorrow's ASPI.

Then these optimum variables were inputted in to the Neural Network model to create a system that will predict the direction of tomorrow's ASPI. Here the past data of the optimum variables was divided in to three sets. The training set consists of the observations for six years except the last six months. Data of the last six months was divided equality in to two sets for testing and validation. Finally, unseen data of three months were given to the trained NN model and Stimulate the Output for Evaluation.

III. Data Description And Data Preprocessing

The variables were used in the study represent the inter market influence, microeconomic factors and news. The US S&P 500 index, Indian BSE Senex index, the Amex oil index, Exchange rate, Interest rate and news were considered as the potential influential variables that affecting the direction of the close prices of the ASPI of day (t+1). The Stock market data was downloaded from www.finance .yahoo.com website and the corresponding symbols for these stock markets are as follows

- All Share price Index - ^CSE
- US S&P 500 - ^GSPC
- BSE Sensex -^BSE
- AMEX Oil Index -^XOI

The macroeconomic data were obtained from the Central Bank of Sri Lanka. Daily news papers were observed to gather 'news'.

The study period was from 1st November 2002 to 31st December 2008. This period covers 1573 daily observations of the respective variable. Therefore, the data set is large enough for modeling and short enough to exclude unnecessary impact from the very past observations. Since different stock markets are close on different holidays, the regular time series data sets considered have missing values. If no trading took place on a particular day, the rate of change of price should be zero. Therefore the missing values were replaced by the corresponding value of the last day. Since the interest was to predict the direction of the close price of the ASPI, the response variable (Y) was introduced as in the Criteria A.

Criteria A

Y = 1 if $RR_{CSE}(t)>0$
Y = 0 Otherwise

Where,

$$RR_{CSE}(t) = \frac{[V_{CSE}(t)-V_{CSE}(t-1)]}{V_{CSE}(t-1)} \quad (1)$$

Here $RR_{CSE}(t)$ and $V_{CSE}(t)$ denotes the Relative Return (RR) and the Close Price of the ASPI on day t, respectively.

Relative Return of the all variables was considered in the study when predicting the close price index of the tomorrow's ASPI.

The qualitative variable NEWS, will be classified the overall situation in Sri Lanka as negative or non-negative news in the sense that the news item of day t negatively affect on ASPI or not, as in the Criteria B,

Criteria B

NEWS = 1 if there was non-negative news on the day t
NEWS = 0 if there was negative news on the day t

The classification of the NEWS was done by considering the factors such as political situation, civil war situation, environment factors, effective price changes of commodities and oil price change in day t.

IV. Genetic Model

In this study GA is employed and created a Genetic Model with the intention of finding the optimum variable combination that will affect the direction of the close price of ASPI of day (t+1). Fig. 1, illustrates the structure of the Genetic model.

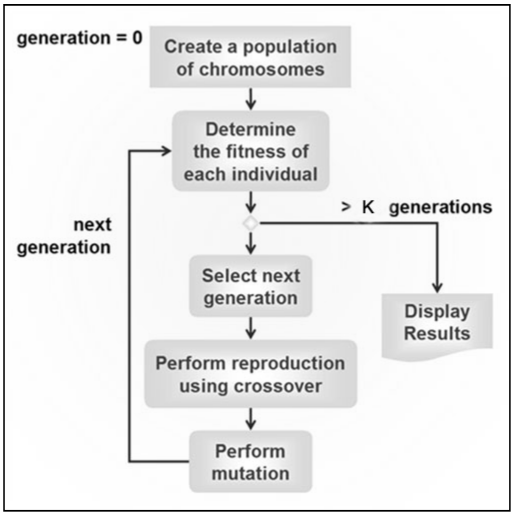


Fig. 1. Structure of the genetic model.

The programming language java was used to implement the model. This Genetic Model was divided into twenty main coding modules before starting the implementation. Final implementation of this Model was carried out by “optimumstockvariables.java” module by integrating the entire main and sub coding modules.

Seven potential input variables were selected for this study. Therefore the number of genes in the chromosome was seven. Table I illustrates the structure of the chromosome. Since the relative return of the each input variable was taken, values of the genes should be binary. So the upper bound should be 1 and the lower bound should be 0.

However, chromosome has no knowledge about the specific problem given by the study, and hence has no intrinsic way of deciding if one potential solution is any better than another potential solution. That's where the fitness function comes in: it's a single method that must implement, which accepts a potential problem solution and returns an integer value that indicates how good (or "fit") that solution is relative to other possible solutions. The higher the number, the better the solution. The lower the number (one being the lowest legal fitness value), the poorer the solution. In this study GA model was used these fitness measurements to evolve the population of solutions toward a more optimal set of solutions. Equation (2), illustrates the fitness function of the model.

$$f(C) = \left[\frac{p}{p+n}\right]\left(\frac{p}{P}\right)\left(\frac{1}{n}\right) \qquad (2)$$

p is the number of positive samples the rule covers, n is the number of negative sample the rule covers, P is the total number of positive samples and C is the rule or the condition. The rule C of the study was, whether the close price of the direction of tomorrow’s ASPI will go up for a particular input combination.

The termination criterion chosen for the Genetic Model is related with the number of generations formed. Thus when the algorithm reaches a specific number of generations it stops and returns the chromosomes structure of the fittest individuals. Max allowed evaluations for the GA model [3] can be changed, by changing the value of the “MAX_ALLOWED_EVOLUTIONS“ variable in the “OptimumStockVariables.java” module. After several changes of the value of the “MAX_ALLOWED_EVOLUTIONS“ variable, it has been decided that the two hundred evaluations are high enough for successful convergence. So by adding further high value for number of evaluations will only add more computational time. Fig. 3, illustrates the result of the Genetic model.

TABLE I
CHROMOSOME OF THE GENETIC MODEL

CSE 1^{st} Lag	GSPC	BSE	EX_R	NEWS	IN_R	XOI

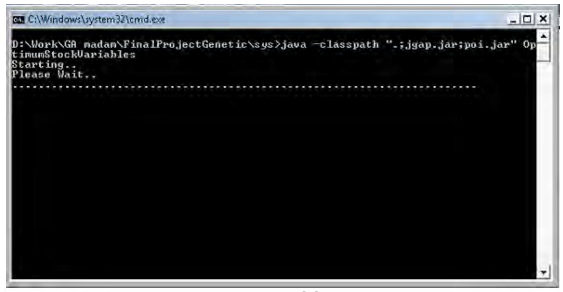


Fig. 2. Genetic Model Starting.

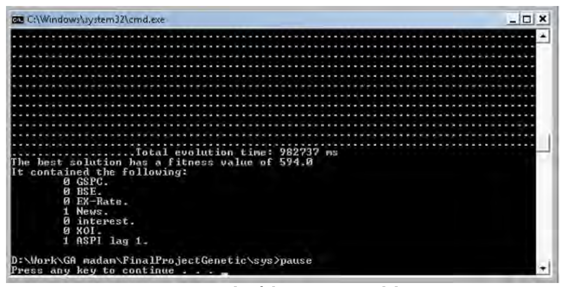


Fig. 3. Result of the Genetic Model

It emphasis that the optimum input variable combination that affects the direction (up or down) of the close price of the tomorrow’s (day (t+1)) ASPI is ‘NEWS’ and ‘ASPI Lag1’. ‘ASPI Lag 1’ stands for the close price index of ASPI of day t and ‘NEWS’ Represents the environment, political and war related information of the day t.

V. NEURAL NETWORK MODEL

Neural network (NN) model was employed to predict the direction of the close price index of the tomorrow’s ASPI value using the optimum input variable combination given by Genetic Model.

The data set considered for the study contains data over six years. But sometimes the earliest time series data did not affects the current time series data. Therefore the data set was divided in to two time windows. So the first window was, from 4th November 2002 to 24th February 2005 and the second window was, from 25th February 2005 to 31st December 2008. Hence two identical neural network models were employed for the predictions; namely NN Model A and NN model B. Past data related to optimum variable combination contains 1572 samples. Those data was divided in to two for the two NN models. Since there should be training and validation data sets to NN to perform well, two samples from those data sets each have hundred observations) were separated for validation and testing.

There were two layers in the neural network model and the weight adjustments did in batch mode [4]. The training algorithm was back propagation algorithm and the error function was mean square error function [5]. Both neural network models (A&B) had these same properties. The Fig. 4, will illustrate NN model structure in a graphical way.

Neural Network tool box of MATLAB R2007a software was used to create the NN models. Data related to training, testing and validation were fed into the models and the outputs of the models were shown in the Fig. 5 and 6.

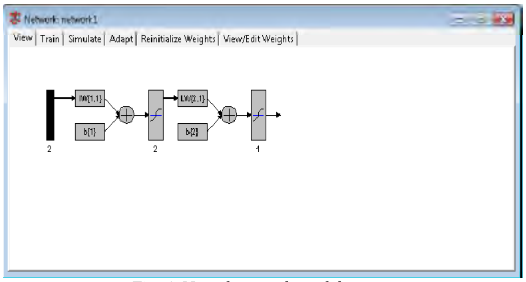


Fig. 4. Neural network model structure

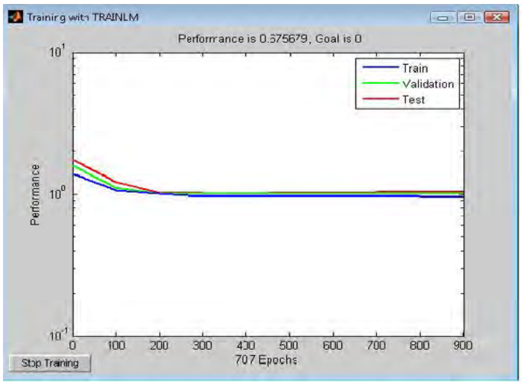


Fig. 5. Training output of the NN Model A

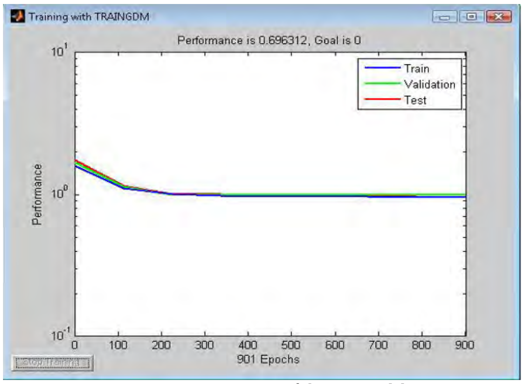


Fig. 6. Training output of the NN Model B

By observing Fig. 5 and 6, it is clear that the NN Model B gives a high performance than NN Model A (since same criteria was used for the both NN Models, normalization of the performance values was not necessary for model comparison).

After Models were created successfully, another unseen data of three months were given to the network models to evaluate the model. Output of this data were recorded using the stimulate option of the NN tool box of MATLAB.

For evaluation, unseen data of three months will be given to the NN model and stimulate the output. Since the transfer function called tansigmoid [4] was used output values will be within the interval [-1 1]. So following criteria was employed to normalize those values.

Criteria C

Y = 1 if *output* > 0
Y = 0 Otherwise

Where,

Y will represent the direction of the close price value of ASPI of day (t+1).
Y=1 represent going up
Y=0 represent going down.
output will be the output given by NN Models.

Then those values were compared with the real ASPI relative return values of the day (t+1) and calculated the percentage of success. The Table II will illustrate the results.

According to the Table II values, it is clear that NN Model B was more successful in predicting than NN Model A. But there is only 2.5% difference between two models so one can argue that the successfulness is nearly same. Also the fact that both models predict the negative results well rather than the positive results cannot be neglected. Positive tab of Table II indicates the percentage number of going ups (Relative Return of ASPI =1) that the NN Model predicts successfully and Negative tab indicates the percentage number of going downs (Relative Return of ASPI =0) that the NN Model predicts successfully.

TABLE II
HIT RATES FOR THE TWO MODULES

	Positive	Negative	Total
NN Model A	58%	63%	60.5%
NN Model B	61%	64%	62.5%
Both Models	59.5%	63.5%	61.5%

VI. CONCLUSIONS

Results of the Genetic model reveal that the best set of input variable combination for predicting the direction of the close price of the ASPI of day (t+1) is news and close price of the ASPI of day t. This study also highlighted that the considered foreign stock markets, US S&P 500 and Indian BSE did not have significant impact on Sri Lankan Stock Market. Amex oil index and exchange rate are also not significant according to the Genetic model. But interest rates are somewhat significant and affects negatively on the close price of the ASPI of day (t+1).

Additionally the results of the study provide evidence to believe the fact that time series data can be somewhat period oriented. By considering all the evidence it is clear that the news and the news related to civil war has a high negative impact on the close price of the Sri Lankan stock market.

Finally, the combination of genetic algorithm and neural network can be employed to create a system that can predict the direction of tomorrow's (day (t+1)) close price of ASPI to a satisfactory extent and this system can be used for taking investment decisions in Sri Lankan Stock Exchange.

A. Comparison with Other Similar Studies

According to Reference [6] overall hit rate of the modified Neural Network approach for predicting the close price index of ASPI is 59.2%. According to [1] the best prediction accuracy for predicting Japan Nikkkei225 index is 68%, US S&P 500 is 63% and UK FTSE100 is 61%. In this study the hit rate is 61.5%. Also it is widely accepted that the ASPI is highly volatile and hence prediction is little bit difficult. Therefore the outcome of this study is good enough when comparing with the other similar studies.

B. Limitations and Improvements Suggested

This study was only limited to two microeconomic variables: the exchange rate and the interest rate. There are other macroeconomic variables which may have an impact on the CSE. And there may be other factors which may directly or indirectly influence the direction of the close price index of the CSE. Due to the time limitations and the difficulties faced in acquiring data those factors were not considered. But if one can include these factors the outcome might be more accurate.

This study incorporated news which is a qualitative representation of the Sri Lankan political situation, civil war situation, environment factors, effective price change of commodities and oil price changes. The information was gathered from the past daily news papers. Hence the direct impact of the news on the Sri Lankan stock market was untraceable. This was a disadvantage to the study.

Furthermore the news published on the news papers is not the only news which has an impact on the stock markets. The behavior of the brokers' manipulation, insider trading, management and policy changes of the companies and many other reasons create news, which affect the stock market trading. Such information is unavailable to the general public.

The genetic model or the system that was developed in this study to identify the optimum input variable combination is a console application. But if this system may goes as a commercial system user may be liked to have a graphical user interface rather than a console interface. So it is good that one can improve this model by adding a graphical user interface.

This conclusion is based on the considered study period and the input variables. However these results may differ according to the nature of the market, the study period and also the input variables.

REFERENCES

[1] M. L. Leung, H. Daouk, and A. S. Chen, "Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models," *International Journal of Forecasting,* 16, pp. 173-190, 2000.

[2] A. Wijayanayake, and T. Rupasinghe, *Identifying the Factors That Affect the Sri Lankan Stock Market*, Proceedings of the International Sri Lankan Statistical Conference: Visions of Futuristic Methodologies, 2004, pp. 381-388.

[3] E. D. Goldberg, *Genetic Algorithm in Search, Optimization, and Machine Learning,* New York: Addison-Wesley, 1989.

[4] M. C. Bishop, *Neural Networks for Pattern Recognition*, New York: Oxford University Press, 1996.

[5] K. Gurney, *An Introduction to Neural Networks*, London: UCL Press, 1997.

[6] C. D. Tilakaratne, J. H. D. S. P. Tissera and M. A. Mammadov, *Predicting Trading Signals of All Share Price Index Using a Modified Neural Network Algorithm*, Proceedings of 9th International Information Technology Conference, Colombo, Sri Lanka, 2008.

An Automatic Measurement Algorithm for the Diameters of Carbon Nanotubes by Using Image Processing

Yoichiro Iwasaki[1], Toshiyuki Nakamiya[1], Ryosuke Kozai[1], Fumiaki Mitsugi[2], and Tomoaki Ikegami[2]

[1]Department of Electronics and Intelligent Systems Engineering, Tokai University

9-1-1, Toroku, Kumamoto 862-8652, Japan

[2]Graduate School of Science and Technology, Kumamoto University

2-39-1, Kurokami, Kumamoto 860-8555, Japan

Abstract- **We propose an automatic measurement algorithm for the diameters of carbon nanotubes (CNTs) electron microscopy images by using image processing. In the algorithm, Otsu's method (discriminant analysis method) was employed to determine automatically the threshold for image binarization. Proposed algorithm provides the numerous values of diameters detected by horizontal and vertical scans for the binary images. The diameter measurements using developed program were carried out for the multi-walled CNTs (MWNTs) taken by a scanning electron microscopy (SEM).**

We have confirmed that the proposed scan algorithm detects well the parts of diameters in the wide measurement areas. We have also confirmed that the measured diameters ware close to the values estimated manually. So, we can say that the developed program is able to detect the diameters for the clearly observed CNTs.

We suppose that this method is useful to evaluate the diameters of the CNTs not only for the management of the CNTs products but also for the optimization of CNTs preparation processes.

I. INTRODUCTION

Since the discovery of CNTs [1], numerous researchers have investigated on single-walled CNTs (SWNTs), multi-walled CNTs (MWNTs), their growth methods, their diameters, their lengths, alignment mechanism, electron emission property, gas sensing property, and device applications.

It is very important to evaluate uniformity of diameters of CNTs in order to increase the uniformity of the characteristics of CNTs for the above mentioned applications. The diameters of CNTs are generally estimate manually using the image taken by a scanning electron microscopy or a transmission electron microscopy (TEM). These human visual works are extremely inefficient.

There is a method using radial breathing mode (RBM) of resonance Raman spectroscopy to estimate automatically the diameters of CNTs [2]. However, this method is effective only for SWNTs and is no use to MWNTs, and estimates only the average diameter in the region of the Raman laser irradiation. Therefore, this method cannot provide the variations of numerous diameters, and also cannot estimate the uniformity of diameters in a wide measurement area.

Recently, a few methods to measure automatically the characteristics of nanomaterials using image processing were proposed.

A measurement method for the characteristics of MWCNTs was proposed [3]. This method provides outer radius, inner radius, and a physical characteristic of MWCNTs. However, the selected areas to measure MWCNTs are restricted by the image segmentation procedures. Therefore, it is not effective to evaluate the uniformity of the diameters of CNTs in a wide area. Other image analysis method for nanocomposites contained CNTs was proposed [4]. However, this method cannot be used to measure CNTs diameters.

In this paper, we propose a diameter measurement algorithm for CNTs electron microscopy images by using image processing. In the algorithm, Otsu's method (discriminant analysis method) [5] was employed to determine the threshold for image binarization. Proposed algorithm provides the numerous values of diameters detected by horizontal and vertical scans for the binary images. The diameter measurement was carried out for the MWCNTs SEM images by the developed program.

II. PROPOSED ALGORITHM TO MEASURE THE DIAMETERS OF CNTS

Fig. 1 shows a MWCNTs image taken by a SEM. The proposed program measures the diameters of CNTs in the rectangle surrounded by four points P_0-P_3. The mixture area of image and characters is not measured because each pixel cannot be distinguished between image and character. Before the image processing is started, the white color frame is added for the circumference of CNTs image. The preparations are necessary to begin the detection for CNTs regions.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_48, © Springer Science+Business Media B.V. 2010

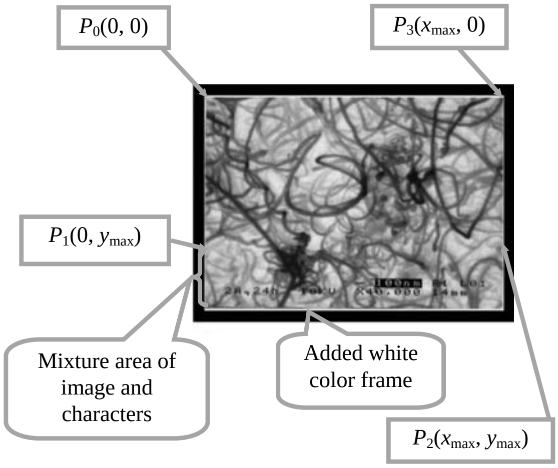


Fig 1. A sample of CNTs image.

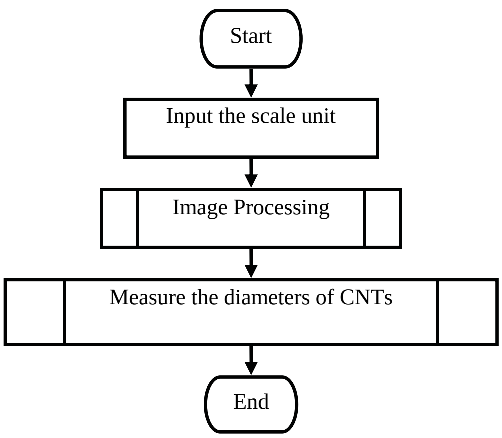


Fig 2. Main program of the diameter measurement.

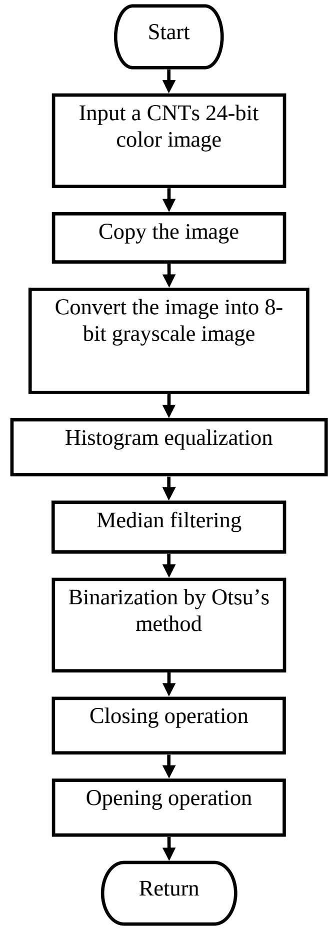
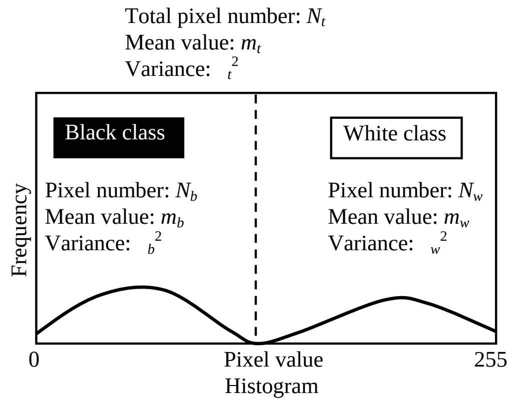


Fig 3. Program of the image processing.

The program is mainly composed from the image processing and the measurement of the diameters of the CNTs as shown in Fig. 2.

At first, the captured CNTs image is treated the noise reduction and binarization by the image processing. Fig.3 exhibits the flow chart and the explanation of the image processing. The captured image is converted to 8-bit grayscale image, and the histogram equalization is carried out to emphasize the contrast. Then, the noise is reduced by the median filter. After that, the automatic binarization is performed using Otsu's method. In the Otsu's method, within and between class variances for black and white classes are used to determine the threshold of the pixel value as shown in Fig 4. We have confirmed that the Otsu's method is effective for the binarization of CNTs images. The closing and opening operations are followed to remove the binarization noise.

After the binarization of the CNTs image, the measurement of diameters of the CNTs is performed. The flow chart of the measurement program is shown in Fig. 5. In this program, x(horizontal) and y(vertical) direction scans are executed to search CNTs, and measure the diameters. Before the scan, the lower and upper limits for the diameter should be set. The procedure how to measure the diameters is explained in Fig.6 using x-direction scan.



Within-class variance: $\sigma^2_{\text{within}} = \dfrac{N_b\sigma_b^2 + N_w\sigma_w^2}{N_b + N_w}$

Between-class variance:

$$\sigma^2_{\text{between}} = \frac{N_b\left(m_b - m_t\right)^2 + N_w\left(m_w - m_t\right)^2}{N_b + N_w}$$

Separation metrics: $\dfrac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}}$

Threshold is given at the maximum of separation metrics.

Fig. 4 The algorithm of Otsu's method.

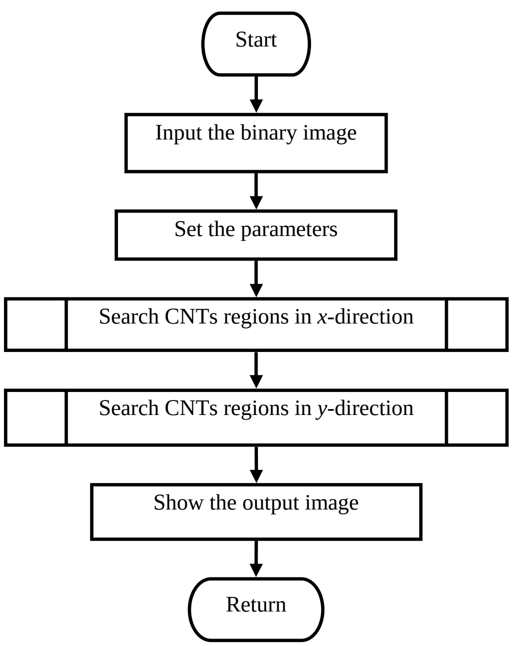


Fig.5 Flow chart of measuring diameters of CNTs.

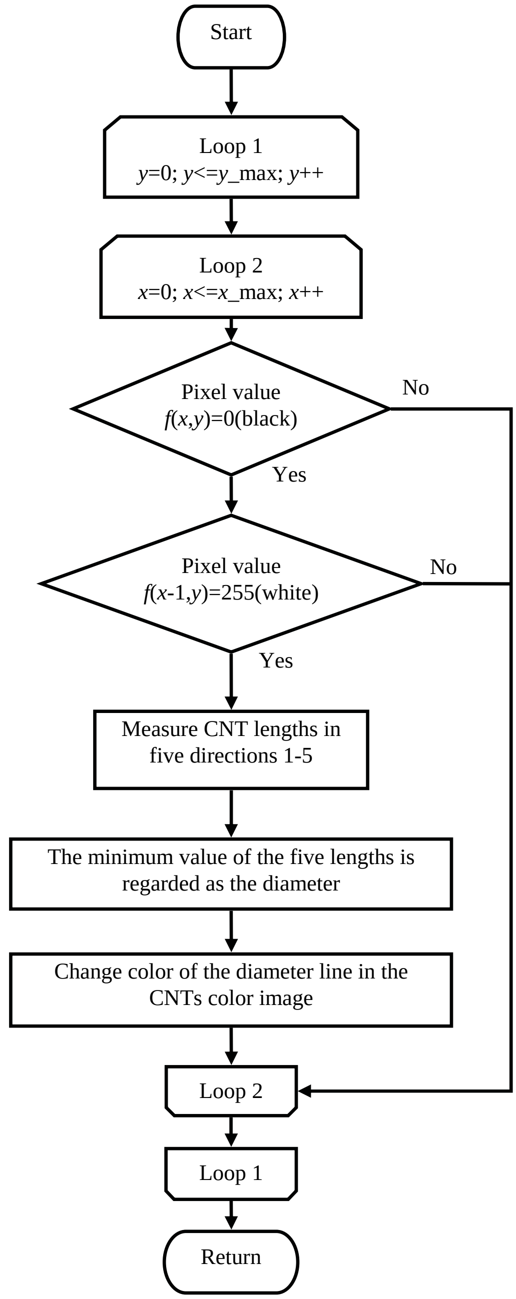
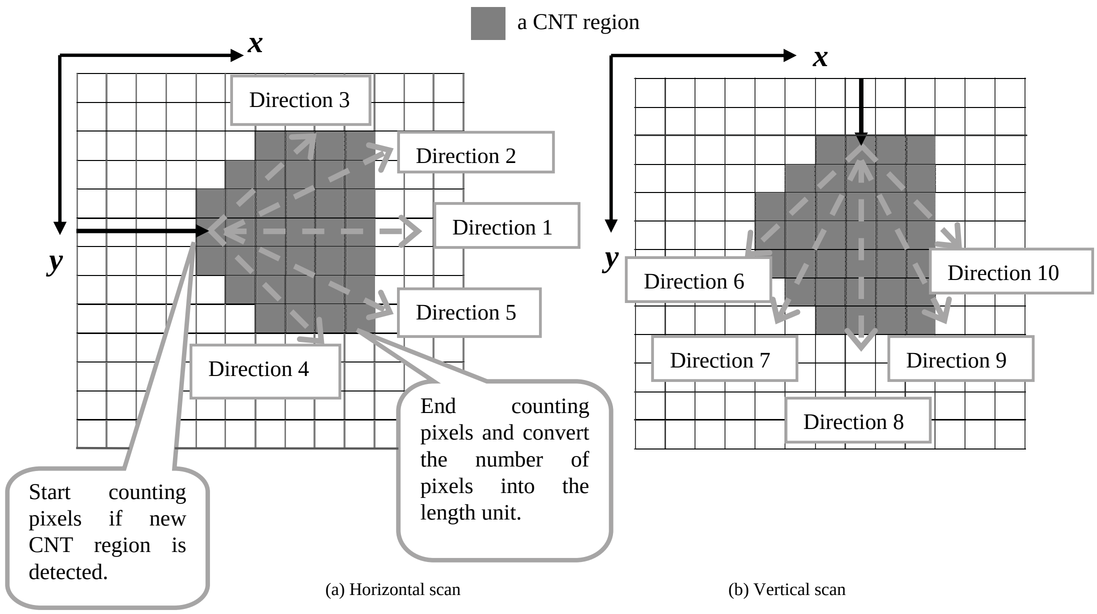


Fig.6 Flow chart of searching CNTs regions in x-direction.



(a) Horizontal scan (b) Vertical scan

Fig. 7 Explanation of the procedure for searching CNTs regions in *x*-direction (horizontal scan) and *y*-direction (vertical scan).

At first, the *y* axis is fixed, and *x*-direction scan starts. If the pixel value at *x* is 0(black) and the pixel value at the *x*-1 is 255(white), the program recognizes that the *x* position is the edge of a CNT, and starts counting black pixels in 5 directions of 0, 25, 45, 315, 335 degrees shown in Fig. 7(a). The minimum length in the 5 measured lengths is regarded as the diameter. The color of the pixels regarded as the diameter are changed, and finally the convert from the number of pixels to its real length is carried out. This process is repeated by increasing *y* positions.

After *x*-direction scan, the scan direction is changed, and *y*-direction scan is executed like the horizontal scan. In vertical scan, 5 directions of 225, 245, 270, 295, 315 are used as shown in Fig. 7(b).

Finally, the output image which indicates the detected diameters is shown.

III. Experimental Results

We have developed our algorithm with Visual C++ 2008 and Computer Vision Library OpenCV [6].

The program was utilized to images of two samples.

After the start of the image processing, the scale unit should be inputted as shown in Fig. 2. The both ends of the scale bar are pointed by a mouse, and the unit value is inputted as shown in Fig.8. The input data are used to convert the numbers of pixels into the real lengths.

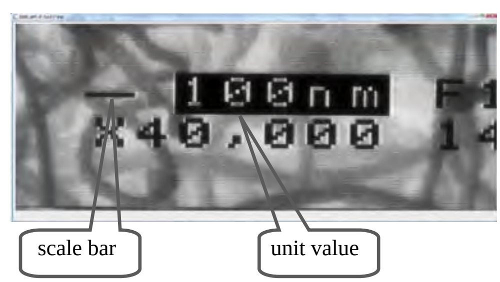


Fig. 8 Scale area in a CNTs SEM image.

In this experiments, we decided that the scanned vertical length between P_0 and P_1 in Fig. 1 was set to 70 % of the height of the image.

Table I shows the summary of results of diameter measurements which include information such as number of succeeded diameter detection, the mean values of diameters, and the standard deviations of diameters.

TABLE I
SUMMARY OF RESULTS OF DIAMETER MEASUREMENTS

		Sample 1	Sample 2
Number of detected diameters	*x*-direction	31,287	12,671
	y-direction	47,651	18,628
	total	78,938	31,299
Mean values of diameters	*x*-direction	20.8nm	23.6nm
	y-direction	19.5nm	23.7nm
	total	20.0nm	23.7nm
Standard deviations of diameters	*x*-direction	11.6	12.9
	y-direction	11.9	12.6
	total	11.8	12.7
width x height of image (pixels)		4,440 x 3,200	4,400 x 3,200
Threshold of Otsu's binarization		125	123

As shown in Table I, the statistical values of detected diameters in case of *x*-direction and *y*-direction scans are almost same because the CNTs are randomly aligned. Fig.9 shows two samples of CNTs images, and diameter detection results in which detected diameters are shown in gray.

We can see that the proposed scan algorithm detects well the parts of diameters in the wide areas. We have confirmed that the measured values of diameters ware close to the values estimated manually. So, we can say that the developed program is able to detect the diameters for the clearly observed CNTs.

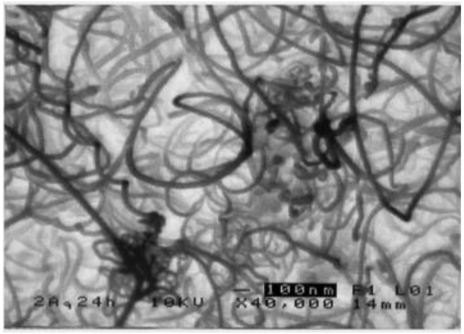
In the experiments, we used a personal computer with an Intel Core i7 2.93GHz CPU and 6.0 GB RAM under Windows Vista 64-bit Operating System, and confirmed that the diameter measurement process was finished within 7 seconds for the sample 1.

IV. CONCLUSIONS

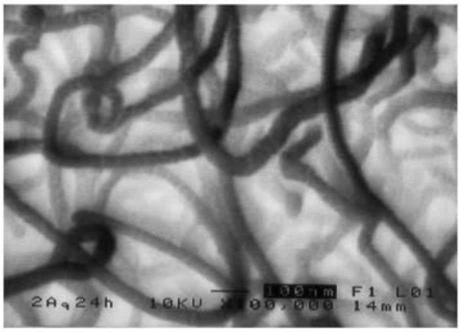
The developed program could detect the diameters of CNTs, and calculate precisely the statistics such as the mean values of diameters, the standard deviations of diameters.

We suppose that this method is useful to evaluate the diameter of the CNTs not only for the management of the CNTs products but also for the optimization of CNTs preparation processes.

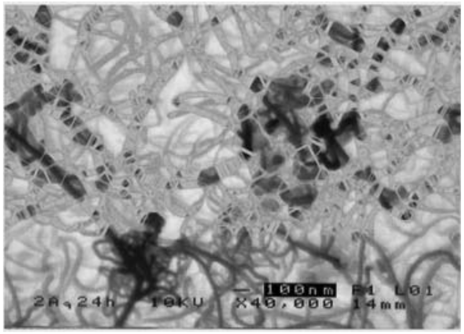
We can obtain more effective information for the depth



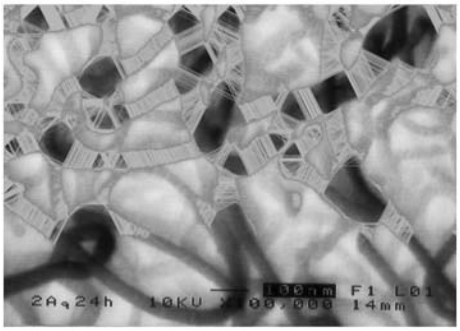
(a) Original image of sample 1



(b) Original image of sample 2



(c) Detected diameters shown in gray of sample 1



(d) Detected diameters shown in gray of sample 2

Fig. 9 Detection results for CNTs diameters.

profile if the depth information is given by other microscopy like 3-dimensional SEM.

REFERENCES

[1] S. Iijima, “Helical Microtubules of Graphitic Carbon," *Nature*, Vol. 354, pp. 56-58, November 1991.

[2] A. Jorio, R. Saito, J. H. Hafner, C. M. Lieber, M. Hunter, T. McClure, G. Dresselhaus, and M. S. Dresselhaus, “Structural (*n*, *m*) Determination of Isolated Single-Wall Carbon Nanotubes by Resonant Raman Scattering," *Physical Review Letters*, Vol. 86, No. 6, pp. 1118-1121, February 2001.

[3] C. Gommes, S. Blacher, K. Masenelli-Varlot, Ch. Bossuot, E. McRae, A. Fonseca, J. -B. Nagy, and J. -P. Pirard, “Image Analysis Characterization of Multi-Walled Carbon Nanotubes,” *Carbon*, Vol. 41, pp. 2561-2572, 2003.

[4] J. Ryszkowska, "Quantitative Image Analysis of Polyurethane/Carbon Nanotube Composite Microstructures," *Materials Characterization*, Vol. 60, pp. 1127-1132, 2009.

[5] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms,” *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 9, No. 1, pp. 62-66, 1979.

[6] G. Bradski and A. Kaehler, *Learning OpenCV*, O'Relly Media, September 2008.

Object Detection in Foveated Images

Abulfazl Yavari[1], H.R. Pourreza[2]
[1]*Payam Nour university of Kashmar, Kashmar, Khorasan Razavi, IRAN*
[2]*Ferdowsi University of Mashad, Mashad, Khorasan Razavi, IRAN*
E-mails: Abulfazl_yavari @yahoo.com, hpourreza@um.ac.ir

Abstract **– In this paper, we present a new visual attention system which is able to detect objects in the images with non-uniform resolution. Since, one of the goals of the visual attention systems is simulating of human perception, and in human visual system the foveated images processed, therefore, visual attention systems should be able to identify the object to these images. We test the system by two types of the images: real world and artificial images. Real world images include the cloth images with some defect, and the system should detect these defects. In artificial images, one element is different from the others only in one feature and the system should detect it.**

Index Terms **– Visual Attention, foveated Images, object detection**

I. Introduction

A. Visual Attention

The visual attention is a selective process that enables a person to act effectively in his complicated environment [1]. We all frequently use the word "attention" in our everyday conversations. A definition for attention is given in [2]: "attention defines the mental ability to select stimuli, responses, and memories". These stimuli can be anything, for instance, think of conditions in which you are studying and concentrated on the text you are reading. Suddenly you are interrupted with a loud sound or you smell something burning and attract your attention. Similarly, there are stimuli in the environment that affect our vision, for example, a moving object, picture on the wall, or a bright area in a dark room. These are examples of cases where without any predetermined goal automatically attract our attention. This type of attention is called bottom-up attention [3](fig. 1).

There are other cases in which we are looking for a special object in the environment, and all the things that have similar features to that object, will attract our attention. Assume, for instance, that we are looking for a red pen on a table. Anything with red color or with a shape like a pen will attract us, and so we may find the desired object in the first or the next focuses. This type of attention is called top-down attention [3] (Fig. 2).

Fig.1 a lamp attract our attention (bottom-up attention)

B. Foveated Images

Human vision system (HVS) is a space variant system [4]. It means that by receding from the gazing point, the resolution gradually decreases and only the totality of the scene will survive. Images that have this feature are called foveated images. The area with the highest resolution is called the fovea [5] (fig.3).

We can find the source of this behaviour, by studying the eye's structure. There are two kinds of vision cells on the retina, cone cells and rod cells. Cone cells are too much less than rod cells in number, but they are sensitive to color, and each cell is individually connected to a nerve. Cone cells are gathered in the fovea area in the retina. Rod cells are too much more than cone cells and they are in the area around the fovea. Multiple rod cells are connected to a single shared nerve and they are sensitive to light [6].

Because humans have non-uniform vision and in fact their brains perform special processing on foveated images, in this paper we concentrate on visual attention on foveated images. Up to now, visual attention is studied only on normal images (uniform resolution).

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_49, © Springer Science+Business Media B.V. 2010

Fig.2 when we looking for a red pen on the table, regions that have red color, attract our attention (top-down attention)

Fig.3 foveated image

Section 2 is dedicated to related works and we introduce what currently is performed in the field of visual attention. In section 3 we present our work with enough details and then the tests and evaluation results are discussed in section 4. Finally we present the conclusions and our future works in section 5.

II. Related Work

The increased interest on research on visual attention together with the increased power of computers and the resulting ability to realize complex computer vision system has led to a wide variety of computational systems on visual attention. In this section we introduce several of the most important computational systems on visual attention.

The first computational architecture of visual attention was introduced by Koch and Ullman [7]. This model provided useful algorithms for later implementations and for many current computational models of visual attention. The idea is that several features are computed in parallel and their conspicuities are collected in a saliency map. A Winner-Take-All network[1] (WTA) determines the most salient region (MSR) in this map, which is finally routed to a central representation. This approach is strongly biological motivated. It is based on the Feature Integration Theory (FIT) [8]. This architecture is merely bottom-up.

One of the earliest implementation of a visual attention system was introduced by Milanese [9], [10]. It is based on the Koch and Ullman model, and uses filter operations for the computation of the feature maps. These model are especially well-suited to be applied to real world scenes since the filter operation provide useful tools for the efficient detection of scene properties contrast or oriented edge. In this model as feature Milanese considers two color opponencies – red-green and blue-yellow --, 16 different orientations, local curvature, and if no color information is available, intensity. To compute the feature specific saliency, Milanese proposes a conspicuity operator which compares the local values of the feature maps to their surround. This operator is motivated from the On-Off and Off-On cells in the cortex and is also a common technique for detecting contrast in images; it is usually referred to as Center-Surround mechanism or Center-Surround difference. The resulting contrasts were collected in so called conspicuity maps, a term that was since then frequently used to denote feature dependent saliency. The conspicuity maps are integrated into the saliency map by a relaxation process that identifies a small number of convex regions of interest. The output of the system is the saliency map that shows a few regions of interest. A drawback of the system is its high computational complexity that results from the many filter operations on different scales.

Another system is Neuromorphic Vision Toolkit (NVT) [10], [11]. The Ideas of the feature maps, the saliency map, the WTA, and the IOR in this system were adopted from the Koch and Ullman model, the approaches of using linear filters for the computation of the features, and conspicuity maps were adopted from Milanese. In this system features are color, orientation, and intensity and all computations are performed on image pyramids. Additionally, in this system use a weighting function for the weighted combination of different feature maps by promoting maps with few peaks and suppressing those with many ones. This technique is computationally much faster than the relaxation process of Milanese and yields good results. The system contain several details that were chosen for efficiency reasons or because they represent a straight forward solution to complex requirement. This approach may lead to some problems and inaccurate results in several cases. For example, the Center-Surround mechanism is realized by the subtraction of different scales of the image pyramid, a method that is fast but very precise. In this system, the color space is RGB that represents colors differently to human perception, which seem not appropriate for a system simulating human behaviour and leads to implausible results, too.

The NVT in its basic version does concentrate on computing bottom-up attention. The need for top-down influences is mentioned but not realised. Navalpakkam and Itti introduce a derivation of their bottom-up model which is able to deal with top-down cues [11]. The idea is to learn feature values of a target from a training image in which the target is indicated by a binary mask. The system learns the feature values from the dif-

[1] A neural network that determines the most salient in a topographical map.

ferent feature maps on different scales. One difficulty with this approach is that it is not clear how bottom-up and top-down cues compete. Since there is evidence that two distinct brain areas are associated with bottom-up and top-down mechanisms in human perception [12], it might be useful to separate the processing also in a computational system.

The next attention system is Hamker. It aims mainly a modeling the visual attention mechanism of the human brain [14]. Hamker's model, shares several aspects with the architecture of NVT: he computes contrasts for several features (intensity, orientation, red-green, blue-yellow) and combines them in feature conspicuity maps. The conspicuity of these maps is combined in a perceptual map that corresponds to the common saliency map.

Another currently considerable system is VOCUS, introduced by Frintrop [1]. It has many similarities to NVT system and it uses feature maps and saliency and IOR, which have been used also in many currently introduced systems. Our proposed system is very similar to this one from the implementation point of view and utilizes many of its ideas. But as one of the goals of vision attention systems is to simulate the human's perception and his vision system, and foveated pictures are used in his vision system, it's important to note that this system is unable to handle foveated images, but our system is capable of locating the goal in such images.

III. SYSTEM DESCRIPTION

In this section we introduce details of our system that try to find region of interest in foveated images. The architecture of system shares the main concepts with the standard models of visual attention, especially based on the model of Koch and Ullman. The implementation is roughly based on the VOCUS (Visual Object detection with a Computational attention System) [1], one of the best known attention systems currently. However, there are several differences in implementation details and type of the input images (Foveated images).

The structure of our visual attention system is shown in fig.4. Before we go into the details of the system, we present a rough overview of the structure, as illustrated in figure; the input of the system is a foveated image, which may be generated by software or hardware. We use software tools to generate images with the fovea in the centre of the image. First, a preprocessing is performed on the input image, which can be performed in two ways: whether by using Gaussian filters with different sizes, so that the larger filters are used in the center of the image and as we move from the center to the periphery, the size of the filter decreases. Another way is to use a 3*3 Gaussian filter and use it multiple times in the center of the image and reducing the number of its use as we go from the center to the periphery.

Next three different feature dimensions are computed: intensity, orientation, and color. For each dimension, the saliencies are computed on different scales and for different feature types, e.g. red, green, blue, and yellow for the feature color. Thereafter the maps are fused step by step, thereby strengthening important aspects and ignoring others.

For each feature we first compute an image pyramid from which we compute scale maps. These present saliencies on different scales for different feature types. The scale maps are fused into feature maps representing different feature type and these again are combined to conspicuity maps, one for each feature. Finally, the conspicuity maps are fused to a single saliency map, with the degree of brightness proportional to the degree of saliency. From the saliency map, the most salient region (MSR) is extracted and the focus of attention (FOA) is directed there.

A. Feature computation

The channel for the feature intensity extracts regions with strong intensity contrasts from the input image. First, the color image is converted into gray-scale. From the gray-scale image, a Gaussian image pyramid with five different scales s0 to s4 is computed. The intensity feature maps is computed by center-surround mechanism extracting intensity differences between image region and their surround, Similar to ganglion cells in the human visual system [12]. The cells are divided into two types: On-center cells respond excitatorily to light at the center and inhibitorily to light at the surround, whereas Off-center cells respond inhibitorily to light at the center and excitatory to light at the surround [12].

According to the human system, we determine two feature types for intensity: the On-center difference responding strongly to bright region on a dark background, and the Off-center difference responding strongly to dark regions on a bright background. Then the maps for each center-surround variation are summed up by across scale addition: first, all maps are resized to scale s2. The maps are added up pixel by pixel. After summing up the scales this yields 2 intensity maps.

Similar, orientation maps are computed from oriented pyramids [10]. The oriented pyramid in fact consists of four pyramids, one for each of the orientation 0, 45, 90,135. The orientations are computed by Gabor filters detecting bar like feature, according to a specified orientation. The orientation scale maps are summed up by across scale addition for each orientation, yielding four orientation feature maps of scales s2, one for each orientation.

To compute the color feature map, the color image is firstly converted into an LAB-image; an LAB image pyramid is generated. From this pyramid, four color pyramids are generated for the distinct colors red, green, blue, and yellow. On these pyramids, the color contrast is computed by center surround mechanism. The maps of each color are rescaling to the scale s2 and summed up into four color feature maps.

B. Fusing saliency

To combine the feature maps, we should decrease the effect of less important maps. To do so, we use the formula

$$X = X / \sqrt{m} \quad (1)$$

In which **X** indicates the map, and **m** is calculated like below: for instance, for the feature of color, we first calculate the number of pixels (Np) with a value more than the threshold t, and then determine the minimum value (min), and **m** for every map is equal to **Np/min**. Finally, formula **1** is used for saliency maps and by adding them pixel to pixel, the saliency map is obtained. Now we should detect the salient areas in the saliency map. This is possible through finding pixels with largest values. These regions are the region that attracts the attention.

We use a green circle to show this region on input images.

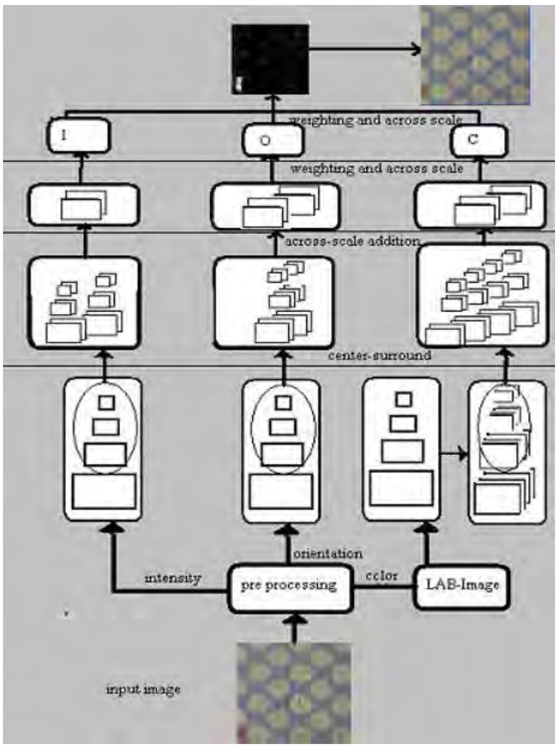


Fig.4 structure of the system

IV. EXPERIMENTS AND RESULTS

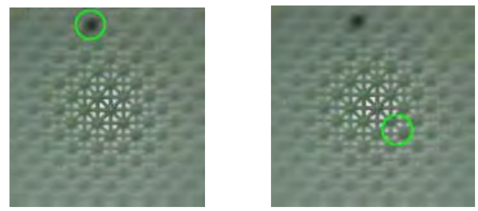
In this section, we present experimental results on real world images and artificial images. We implement our system with MATLAB programming language.

Our sample pictures include real images of cloth and tile with some defects, and the system should detect these defects. In artificial images, one element is different from the others only in one feature and the system should detect it. In each experiment, we first show that VOCUS is unable to find the defects and then show that our system is able to detect the defects correctly. It's worth noting that the small green circle indicates the approximate location of the defect.

A. *Experiment 1*

This image is a tile image that in the top of it there is a black dot (fig.5).

As illustrated in figure, our system detected the defect, but VOCUS not able to detect it. Since the resolution in the center of image is better than periphery, the VOSUS worked wrongly.

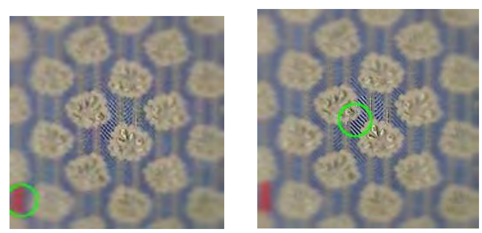


(a) (b)

Fig.5. (a) the defect is detected by our system (b) but not by the VOCUS

B. *Experiment 2*

This image is a cloth image that there is a vertical red defect on it: the defect is detected by our system, but not by the VOCUS (fig.6)

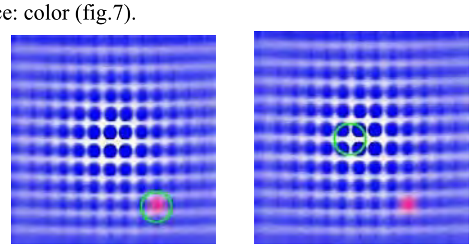


(a) (b)

Fig.6. (a) the defect is detected by our system (b) but not by the VOCUS

C. *Experiment 3*

This image is an artificial image (pop-out). There is a red blob among the blue one. In this image only one feature is difference: color (fig.7).

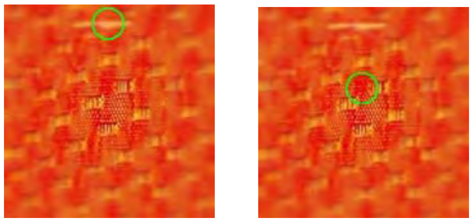


(a) (b)

Fig.7. (a) red bolo is detected by our system (b) but not by the VOCUS

D. *Experiment 4*

This image is also a cloth image, that there is a yellow defect on it (fig.8)



(a) (b)

Fig.8. (a) horizontal line is detected by our system (b) but not by the VOCUS

V. CONCLUSION AND FUTURE WORKS

We presented a new visual attention system which is enabling of detecting attentive areas in the images with Non-uniform resolution, something that no other currently existing system is capable of. According to the standard attention model of Koch and Ullman, and VOCUS, we concentrate on the three features (color, intensity, orientation), since they belong to the basic feature of human perception and are rather easy to compute. Of course, this can only approximate human behaviour where many other features attract attention, e.g., size, curvature, and motion [1].

In future works, we intend to complete the top-down part of the system. In addition, we should improve the system's speed to make it possible for being used in real time cases. We will also study addition of the motion feature, which has an important role in attracting the attention.

ACKNOWLEDGEMENT

The authors would like to thank Simone Frintrop for useful comments on this paper.

REFERENCES

[1] Frintrop, S: VOVUS: a Visual Attention System for Object Detection and Goal Directed Search. PhD thesis University of Bonn Germany. January 9, 2006

[2] Corbetta, M.: Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural system? Proc. Of the National Academy of sciences of the United States of America, 95:831-838. 1990

[3] Desimone, R. and Duncan, J.: Neural mechanism of selective Visual attention. Annual reviews of Neuroscience, 18:193-222. 1995

[4] Coltekin, A.: Foveation for 3D visualization and Stereo Imaging. PhD thesis Helsinki University of Technology Finland. February 3, 2006.

[5] Chang, E.: Foveation Techniques and Scheduling Issues in Thinwire Visualization. PhD thesis New York University. 1998.

[6] Gonzales, R. C. and Woods, R. E.: Digital image processing. Addison-Wesley Publishing Company, 1992

[7] Koch, C. and Ullman, S.:Shifys in selective visual attention: Towards the underlying neural circuitry. Human Neurobiology 4 (4, 1985) 219-227.

[8] Treisman, A. M. and Gelade, G.: A feature integration theory of attention. *Cognitive Psychology* 12 (1980) 97–136.

[9] Milanese, R.: Detecting Salient regions in an Image: From Bilogical Evidence to Computer Implementation. PhD thesis, University of Geneva, Switzerland. 1993.

[10] Milanese, R., Wechsler, H., Gil, S., Bost, J., and Pun, T.: Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In proc. Of the IEEE Conference on Computer Vision and Pattern Recognition, pages 781-785. 1994.

[11] Itti, L., Koch, C. and Niebur, E.: A model of saliency-based Visual Attention for Rapid Scene Analysis. IEEE Trans. On PAMI 20, pages 1254-1259, 1998.

[12] Palmer, S. E.: Vision Science, Photons to Phenomenology the MIT Press 1999.

[13] Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., and Anderson, C.: Overcomplete steerable pyramid filters and rotation invariance. IEEE Computer Vision and Pattern Recognition (CVPR), pages 222-228, 1994.

[14] Hamker, F. Modeling Attention: From computational neuroscience to computer vision. In:proc. Of WAPCV'04 (2004) 59-66.

Digital Semi-Controlled Rectifier using a Microcontroller

Emmanuel E. Rached, Adel I. Chit, Michel J. Owayjan
American University of Science and Technology, Beirut, Lebanon
Department of Computer and Communications Engineering
eer004@students.aust.edu.lb, achit@aust.edu.lb, mowayjan@aust.edu.lb

Abstract- **Controlled rectifiers are an important aspect in power electronics; they are used to modify the average DC voltage at the output. Their use is widespread and they hold an important role in speed control of motors. Traditional analog control for these rectifiers was not always accurate and some required complex designs. This paper presents a new low cost, high accuracy digital control method using a microcontroller. The microcontroller keeps track of the current AC half-wave and triggers the corresponding thyristors based on the desired time delay through the use of a potentiometer. Experimental tests showed a high accuracy and low response time in controlling the thyristor's firing angle.**

I. INTRODUCTION

Most industrial applications use thyristor-based controlled rectifiers, especially where high voltages and currents are involved [1, 2]. Rectifiers are used industrially to provide a variable DC voltage from AC sources to control the speed of motors [3]. They are also used in high voltage converter stations to provide long distance transmissions. Other uses include light dimming, power switching, and certain circuit breaker applications [1, 2]. Traditional analog control of thyristor-based rectifiers not only requires complex designs, but also lacks accuracy [4]. Digital control using a microprocessor can improve the accuracy of such rectifiers that can be used to drive a motor as a load [5].

In this paper we present a new low-cost, high accuracy digital control method using a microcontroller. The proposed solution that was implemented and tested is composed of three sections. The first section monitors the AC supply voltage in order to calculate the zero-crossings of the sine wave. The second part consists of a PIC16F877A microcontroller which calculates and controls the firing angle of the thyristors. And the last stage provides isolation of the microcontroller from the load. The details of the proposed solution are discussed in the next section. Then, we present the tests and the results, and end with some important conclusions and future work.

II. PROPOSED SOLUTION

Controlling the rectifier in a reliable manner requires monitoring of the AC sine wave, the ability to detect the zero crossing, monitoring the potentiometer for the desired firing angle, separating the control circuit from the rectifier and accurately triggering the thyristors. The proposed solution has three main parts. The first part provides the monitoring of the AC sine wave and zero crossing detection. Fig. 1 shows the circuit diagram of this part. The transformer steps down the 220 VAC source to a more manageable 6 VAC voltage. Filter capacitors are used on the both sides of the transformer. The 6 VAC is further reduced to 3 VAC by the use of a voltage divider. The 3 VAC is required so as not to damage the MCP601 operational amplifiers [6]. The sine wave is then fed to two MPC601 to provide a negative and positive reference for the microcontroller.

Unlike the more commonly used 741 Operational Amplifier which requires a 15 V source to operate reliably, the MCP601 can perfectly operate with the same 5 V supply as the microcontroller allowing the use of a single power supply. The lower MCP601 takes as input the sine wave and, in the configuration shown here with the +5 V to V_{dd}, ground to V_{ss} and comparing to ground, outputs a constant 5 V whenever the sine wave is in the positive half-wave. This provides a logical "1" (ON) to the microcontroller so it can reliably know that the signal is currently in the positive half-wave.

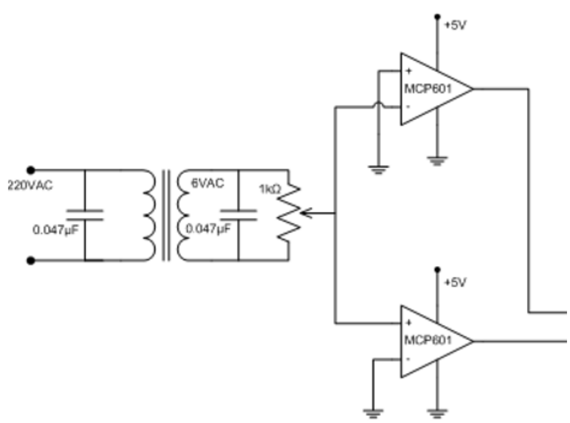


Fig. 1. Sine wave monitoring

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_50,

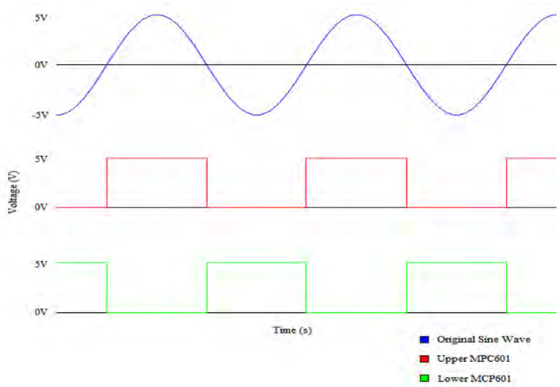


Fig. 2. Operational Amplifiers output

When the sine signal is in the negative half-wave, the Operational Amplifier cannot output a negative voltage since its negative power supply is connected to ground thus making its output a logical "0" (OFF) allowing the microcontroller to know that the signal is no longer in the positive half-wave. The upper MCP601 operates in the same manner except it outputs a signal when the sine wave is in the negative half-wave portion. The output of the MCP601s is a square wave signal with a 5 V amplitude level and 50% duty cycle. Fig. 2 shows the output of both the MCP601s versus the 50 Hz AC input.

The second stage controls the firing angle through the use a microcontroller. A potentiometer connected between +5 V and ground provides the microcontroller with the desired thyristors' firing angle. This connects to the first Analog to Digital converter of the microcontroller. A full +5 V as input instructs the microcontroller to provide a trigger pulse at 0º and 180º for the positive and negative half-waves respectively, effectively turning the controlled rectifier into a normal bridge rectifier. On the other hand, a 0 V input results in the microcontroller outputting a trigger pulse at 180º and 360º for the positive and negative half-waves respectively, effectively negating any output and making it 0 V. The remainder of the range allows precise control of the firing angle between 0º and 180º. The signals coming from the first stage are directed towards pin number 17 for the negative portion and pin number 18 for the positive part allowing the microcontroller to correctly know in what half-wave the sine signal is currently in. The trigger pulse is then outputted on pin 24 for the negative half-wave and on pin 23 for the positive half-wave. Fig. 3 shows the configuration of the microcontroller. The other elements shown in the illustration above are standard configuration for a PIC16F877A microcontroller [7]. The crystal oscillator used is operating at 4 MHz. Fig. 4 shows the potentiometer voltage versus the firing angle.

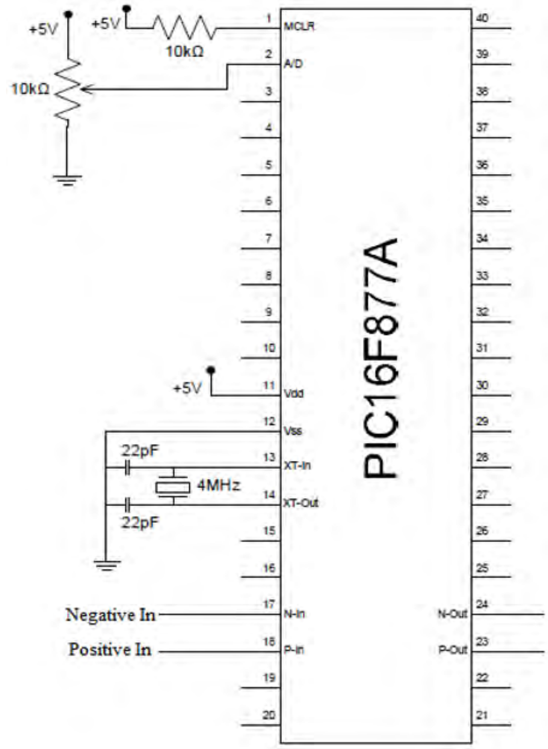


Fig. 3. Microcontroller configuration

As seen from the graph, the firing angle increases from 0º to 180º as the voltage decreases from 5 V to 0 V. This was a design choice. The microcontroller's Analog to Digital converter allows for 10 bits of precision translating into 2^{10} = 1024 possibilities, dividing the 180º into 1024 steps of approximately 0.1758º each. This provides a very high accuracy in controlling the firing angle. The 4 MHz crystal oscillator provides fast execution time for the microcontroller and the software optimizations were performed based on it.

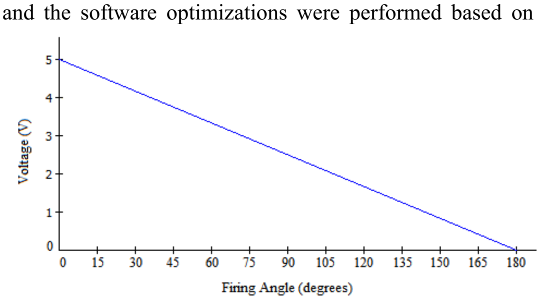


Fig. 4. Firing angle versus Voltage

The following simplified pseudo code summarizes the microcontroller's main BASIC code:

```
Input:   negative_input(boolean)
         positive_input(boolean)
         delay = A/D potentiometer voltage

Output:  negative_output(boolean)
         positive_output(boolean)

While(positive_input = 1)
         {
         delay = GetADValue()
         wait(delay)
         positive_output = ON
         wait(duration)
         positive_output = OFF
         }
```

The negative half-wave's pseudo code is similar to what is shown above.

The third and final part of the solution provides separation between its various components. 1 kΩ resistors placed between the output pins of the microcontroller and the base of the BJTs limit the current draw from the microcontroller to a maximum of 5 mA to prevent damaging it. Pulse transformers provide the necessary separation between the transistors and the thyristors as those will most likely be used in high current circuit. Fig. 5 shows the diagram of this stage. The upper part handles the negative half-wave while the lower one handles the positive half-wave. Freewheeling diodes are used between the components to eliminate flyback.

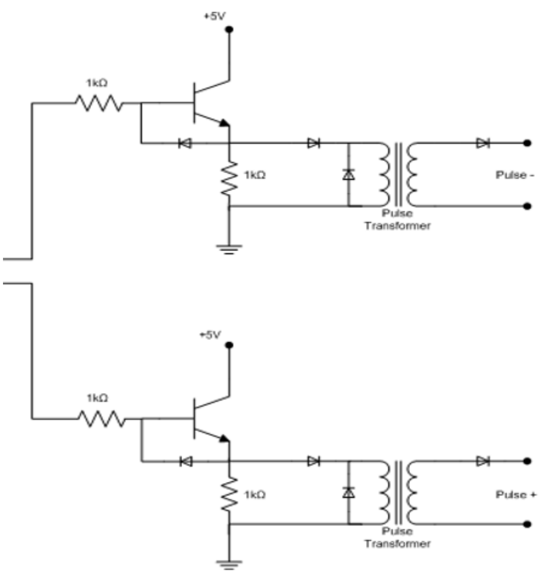


Fig. 5. Final Stage

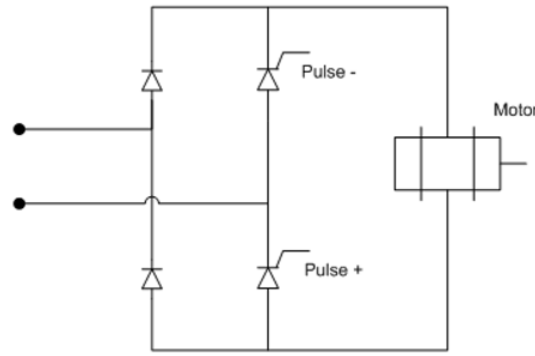


Fig. 6. Semi-controlled rectifier example

In the configuration shown, the transistors' output remains off until a pulse is sent by the microcontroller on the corresponding pin, at which point the transistors switches ON and remains in this state for the duration determined by the microcontroller.

An added simplicity for this solution is that all the active elements in the circuit can be powered from a single +5 V power supply, simplifying power requirements by allowing the use a single 5 V power supply. A standard 5 V, Zener regulated power supply was also implemented to accompany this solution. Fig. 6 shows an example of a semi-controlled bridge rectifier with a motor as the load implementing the solution presented in this paper with the corresponding negative pulse and positive pulse indicated.

III. TESTS & RESULTS

The solution presented in this paper was tested on a 50 Hz, 220 VAC source. A PIC16F887A microcontroller was used as the main controller. The only reason that this particular model was chosen is that it was readily available. A cheaper microcontroller could be used with a slight code modification to achieve cost reductions. A standard 5 V Zener regulated supply was also designed to power all active components of the circuit. Initial testing required some refinement for the code due to delays by the microcontroller in executing the given instructions as different instructions require a different number of cycles to complete. Table 1 shows the results prior to the modifications while Table 2 shows them after the modifications and optimizations were made.

TABLE 1
RESULTS PRIOR TO MODIFICATIONS

Desired Angle	Actual Angle	Difference
0º	3º	~300%
2º	4º	100%
25º	27º	8%
75º	76º	2%
135º	137º	1.5%
180º	188º	4.5%

TABLE 2
RESULTS AFTER MODIFICATIONS

Desired Angle	Actual Angle	Difference
0º	0.036º	~0.3%
2º	2º	0%
25º	25º	0%
75º	75º	0%
135º	135º	0%
180º	181º	0.5%

IV. CONCLUSIONS

This microcontroller powered digital control method is presented in this paper to provide a low cost, high efficiency alternative to currently available methods. Programmed in BASIC, the implementation shows both its high efficiency and low response time compared to other solutions. This presents a potential for cost and time saving for many industries. Some limitations do exist for this particular solution. The code was written for 50 Hz AC electrical sources. Also, the optimizations performed were based on the use of a 4 MHz crystal oscillator. On the other hand, these drawbacks are software based and thus can be tailored to other requirements by simply modifying the microcontroller's BASIC code. Future work could include the use of a more general code for the ability to work with any source's AC frequency and further optimizations. Other improvements would be the use of a cheaper microcontroller as the PIC16F877A was only used since it was immediately available. These improvements are expected to further reduce the cost of this solution and make it more widely available.

REFERENCES

[1] J. Millman and A. Grabel, *Microelectronics*, 2nd ed., Mc-Graw-Hill. 1987.

[2] M. Rashid, *Power Electronics: Circuits, Devices, and Applications* , 3rd ed., Prentice Hall, 2003.

[3] Wikimedia Foundation Inc., "Rectifier." [Online]. 2009. Available: http://en.wikipedia.org/wiki/Rectifier.

[4] D. Maksimovic, R. Zane, and R. Erickson, "Impact of digital control in power electronics," *Proceedings of 2004 International Symposium on Power Semiconductor Devices & ICs*, Kitakyushu, 2003, pp. 13-22.

[5] K. Rahman, M. Choudhury, and T. Zulfikar, "An AT89C51 microcontroller based control circuit for dual three phase controlled rectifier,"*Proceedings of the 3rd International Conference on Electrical & Computer Engineering ICECE 2004*, Dhaka, Bangladesh, December 2004, pp. 347-350.

[6] Microchip Technology Inc., *Datasheet of MCP601/2/3/4*, [Online]. Last accessed October 2009. Available: http://www.microchip.com. 2004.

[7] Microchip Technology Inc., *PIC16F87XA Datasheet*, [Online]. Last accessed October 2009. Available: http://www.microchip.com. 2003.

Generalized Minimum Variance Controller with Dynamic Pole Assignment to Improve Performance in Industrial Applications

M.A. Paz Ramos[1], S.C. Garibo Esquivel[2], O.J. Santos Sánchez[3], R. Fernández del Busto[4], J.F. Briones de la Torre[1], J.L. Gallegos Ramirez[1], J. I. Rodríguez Martínez[1]

[1] Universidad Politécnica de Aguascalientes, Aguascalientes, Aguascalientes, México.
[2]Instituto Tecnológico de Aguascalientes, Aguascalientes, Aguascalientes, México.
[3]Universidad Autónoma del Estado de Hidalgo, Pachuca Hidalgo, México.
[4]Tecnológico de Monterrey Campus Ciudad de México, México Distrito Federal, México.

***Abstract* – A comparison is made between the Generalized Minimum Variance Controller (GMVC) and a new version of it, the Generalized Minimum Variance Controller with a Dynamic Pole Assignment (GMVCDPA) regarding stability and performance under process parameter variations. Simulation examples show that the proposed structure has better characteristics than the original GMVC. To further illustrate the advances of the GMVCDPA, three laboratory experiments are carried out: for DC motor, speed and position control and for chamber prototype, temperature control. Results obtained in temperature control experiment are compared with those obtained with an industrial PID controller.**

I. INTRODUCTION

A good deal of control problems found in industry, are properly solved with conventional control strategies; however, there are some areas in which these techniques are not adequate, and where adaptive controllers have showed better results. Particularly, self-tuning regulators (STR) are being used and have become the most popular adaptive applications in the industry.

A particular STR structure is studied in this work: the Generalized Minimum Variance Controller (GMVC), which is described in section II. The design of the cost polynomials for GMVC, according to a pole assignment is discussed in section III. In order to achieve better results in practice, we propose an algorithmic modification of the original GMVC in section III. In section IV, we discuss stability issues and in section V performance issues are revisited. Finally in section VI physical implementations are presented.

II. GENERALIZED MINIMUM VARIANCE CONTROLLER

Clarke and Gawthrop proposed the generalized minimum variance controller (GMVC) [1], defining a pseudo-output of the system $\phi(t)$ which includes not only the system output, but the reference and control signals as well. Among other properties, this algorithm allows the user to tailor the performance of the controller through a specific choice of the optimality criterion, and also deal with non-minimum phase systems. In the GMVC a generalized output, or pseudo output is defined, by the following equation:

$$\phi(t+k)=Py(t+k)+Qu(t)-Rr(t) \tag{1}$$

where P, Q, R (called cost polynomials) are used to modify the control performance [2] while y is the process output, u is the control input, r is the reference signal and k is the system delay. The cost function to be minimized is the mathematical expectation of $\phi(t+k)$,

$$J=E_x[\phi^2(t+k)] \tag{2}$$

The control law proposed for the GMV controller is

$$Hu(t)+Gy(t)+Er(t)=0 \tag{3}$$

where the control polynomials H, G and E are directly estimated from (1), because in an implicit way, the following relationship prevails:

$$PC=FA+z^{-k}G \tag{4}$$

where:

$$H=BF-QC \quad \text{y} \quad E=-CR \tag{5}$$

In Figure 1, we can see a synthesis of the structure of the GMVC where A, B and C are the process polynomials. The control algorithm is clearly explained in [2].

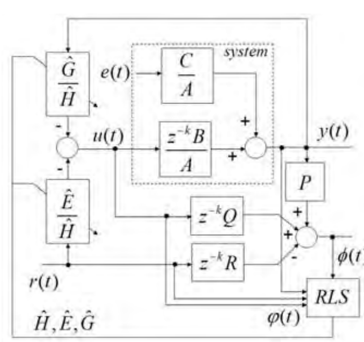


Figure 1. Functional diagram of a GMVC

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_51, © Springer Science+Business Media B.V. 2010

III. GENERALIZED MINIMUM VARIANCE CONTROLLER USING DYNAMIC POLE ASSIGNMENT

The GMVC is a very useful controller when the process is time invariant. Another algorithm related to GMVC is the Generalized Minimum Variance Controller with Pole Assignment (GMVCPA) [2], where the main purpose is to tailor the transient response by pole assignment. For the GMVCPA, the cost polynomials are designed for the controller to achieve a pole assignment, thus solving the identity:

$$PB + QA = T \tag{6}$$

where T is the pole assignment polynomial. Equation (7) is a conceptual representation of the closed loop using a GMVPA control, when the cost polynomials are designed from (6).

$$G_{CL} = \frac{BQ}{PB + QA} \tag{7}$$

The cost polynomials P and Q are fixed during control, and then, if the process parameters are perturbed, the relationship (6) could be invalid. Even when the GMVCPA preserve the ability to minimize the output variance, the performance could be affected due to changes in the denominator of (7). The Generalized Minimum Variance Controller with Dynamic Pole Assignment (GMVCDPA) was proposed in [3], and allows a better control when parameter perturbations are present. While in the GMVCPA the cost functions are proposed off line, in the GMVCDPA the cost polynomials are calculated and updated on line. If we wanted to achieve a dynamic pole assignment in the traditional GMVCPA, we would need two parametric estimations: process estimation and control estimation. In [3] the control estimation is avoided and substituted by a control calculation using only process estimation. Figure 2 illustrates the structure of the GMVCDPA. In this paper, we present a stability analysis and performance test in order to compare the original formulation of GMVCPA against its modification, the GMVCDPA.

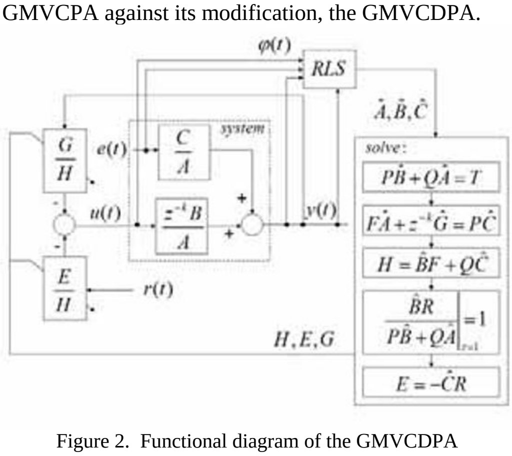


Figure 2. Functional diagram of the GMVCDPA

IV. IMPROVING STABILITY

Second order systems can be considered as representative case in industrial process control. A common recommendation during implementation of self tuning regulators is to include filters to achieve a *relevant identification* [4]. In order to illustrate how GMVCDPA improves stability and performance, we choose a second order system. Although the structure of the GMVCDPA is discrete when it is applied, in this paper we use a continuous analysis, because it makes it easy to visualize the stability zones and, consequently it is easier to appreciate the improvement of the original structure. Consider the next process:

$$\frac{B(s)}{A(s)} = \frac{K}{s^2 + a_1 s + a_2} \tag{8}$$

then n_A=2 and n_B=0, where n is the system order and the subscript refers to the polynomial, we design a GMVCPA controller where n_T=n, if n_P= 1 and n_Q=1, then P and Q could be expressed as:

$$\begin{aligned} P(s) &= p_0 s + p_1 \\ Q(s) &= q_0 s + q_1 \end{aligned} \tag{9}$$

For the next analysis, we assume that the process parameters are the true ones. Considering (6) and (9), when n_T=2, q_0=0 then we have

$$T(s) = s^2 q_1 + s(p_0 K + a_1 q_1) + p_1 K + a_2 q_1 \tag{10}$$

If T(s) is monic, q_1=1.

Applying the Routh criterion to the polynomial (10), we obtain the array

$$\begin{array}{lll} s^2 & 1 & p_1 K + a_2 \\ s^1 & p_0 K + a_1 & \\ s^0 & p_1 K + a_2 & \end{array} \tag{11}$$

There are necessary and sufficient conditions for closed loop stability if

$$\begin{aligned} p_o K + a_1 &> 0 \\ p_1 K + a_2 &> 0 \end{aligned} \tag{12}$$

Consequently, the stability boundaries for possible variations in the process parameters are defined by

$$\begin{aligned} \partial a_1 &> -p_o K \\ \partial a_2 &> -p_1 K \end{aligned} \tag{13}$$

Thanks to all the evidence that has been presented through of this section, we confidently postulate the following Theorem:

Theorem 1 Given the open loop process

$$G(s) = \frac{K}{s^2 + a_1 s + a_2} \tag{14}$$

and a GMVPA controller whose cost polynomials are defined by

$$P(s) = p_0 s + p_1 \qquad Q(s) = 1 \tag{15}$$

assuming a K fixed in (14), the stability zone before perturbations in the process parameters a_1 and a_2 when the loop is closed considering (15), is the S zone whose boundaries are defined by

$$\partial a_1 > -p_o K \qquad \partial a_2 > -p_1 K \tag{16}$$

In the parameters space a_1-a_2 (see Figure 3).

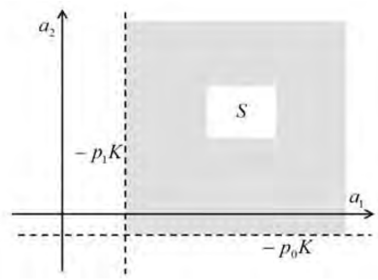


Figure 3. Stability zone, defined by the cost polynomial P

In order to exemplify the **Theorem 1**, we choose to control the capacitor voltage of an RLC circuit which is the low pass filter whose transfer function is

$$\frac{Eo(s)}{Ei(s)} = \frac{1/LC}{s^2 + R/Ls + 1/LC} \tag{17}$$

The values of the electric components are L=8.4572H, C=0.004523F and R=69.02119Ω. Then the transfer function can be rewritten as

$$\frac{Eo(s)}{Ei(s)} = \frac{26.14}{s^2 + 8.161s + 26.14} \tag{18}$$

If the desired T polynomial is defined as

$$T(s) = s^2 + 3.68s + 9.688 \tag{19}$$

to achieve a t_s=2.5 sec. with a M_p=10%, then, when GMVCPA is considered, the cost polynomials are

$$P(s) = -0.17141579\,s - 0.62941652 \qquad Q(s) = 1 \tag{20}$$

then the stability boundaries are

$$\partial a_1 = 4.48 \qquad \partial a_2 = 16.45 \tag{21}$$

For a simulation we calculate the discrete solution for (6) as

$$P(z) = -2.597z + 1.7988 \qquad Q(z) = 1.25784 \tag{22}$$

In Figure 4(a) the output of the process controlled is showed while in the Figure 4(b) we can see the control signal. Suppose now, that there is a perturbation in one of the parameters, particularly in the resistance one. The nominal value is drastically perturbed to R=37.5 in the 25th second. Incorporating the change to (17), the new process coefficients are a_1*=4.434 and a_2*=26.14, which causes for the coordinate (a_1*, $a2$*) to stay outside of the stability zone S. This situation can be appreciated thanks to the simulation results showed in Figure 5.

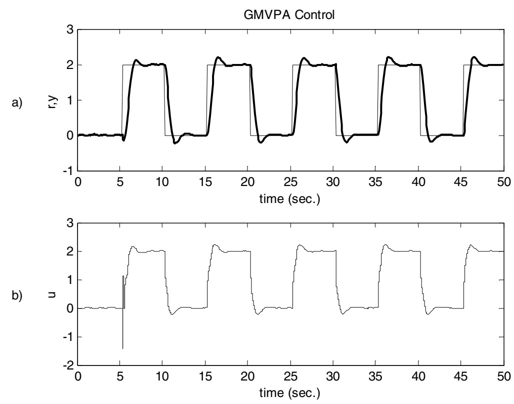


Figure 4. GMVCPA for the RLC circuit example

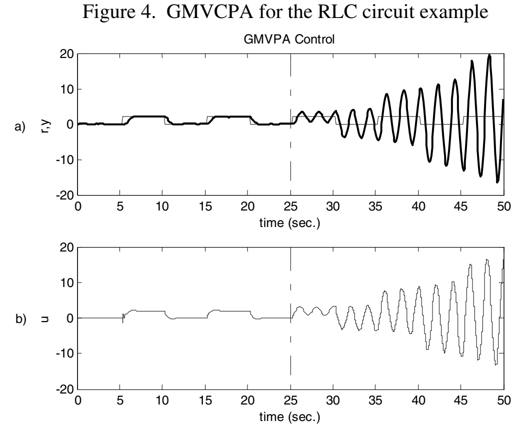


Figure 5. GMVCPA when there is a parameter perturbation

Considering now the parametric uncertainties treated in Example 1, we present the polynomial

$$A^*(s) = s^2 + (a_1 + \Delta a_1)s + (a_2 + \Delta a_2) \tag{23}$$

from (23) we define

$$a_1{*} = a_1 + \Delta a_1 \qquad a_2{*} = a_2 + \Delta a_2 \tag{24}$$

then the closed loop polynomial is

$$T(s) = s^2 + s(p_0 K + a_1{*}) + p_1 K + a_2{*} \tag{25}$$

In the GMVCPA the cost polynomials are designed off line, from the solution of a Sylvester matrix [2].

$$\begin{bmatrix} q_1 \\ q_0 \\ p_1 \\ p_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{K}(t_2 - a_2) \\ \frac{1}{K}(t_1 - a_1) \end{bmatrix} \tag{26}$$

where t_1 and t_2 are the parameters of the T polynomial.

If we consider that the cost polynomial are designed with the nominal parameters a_1 and a_2, coordinate (a_1*, $a2$*) could be out of the stability zone S. On the other hand, when we apply the GMVCDPA, the calculation of the cost polynomials was performed on line using the estimation. Assuming the asymptotic convergence of the estimation we have

$$\begin{aligned} \hat{a}_1 &\approx a_1 + \Delta a_1 \\ \hat{a}_2 &\approx a_2 + \Delta a_2 \end{aligned} \tag{27}$$

then the calculus of the cost polynomials is defined by

$$\begin{bmatrix} q_1 \\ q_0 \\ p_1 \\ p_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{K}(t_2 - \hat{a}_2) \\ \frac{1}{K}(t_1 - \hat{a}_1) \end{bmatrix} \tag{28}$$

if we substitute (28) in (16)

$$\begin{aligned} \partial a_1 &= a_1 * - t_1 \\ \partial a_2 &= a_2 * - t_2 \end{aligned} \tag{29}$$

using (29), we can see that if T(s) fulfils the Stodola condition, the boundaries of the stability wrap the perturbed coordinate (a_1*,a_2 *).

From the previous results outlined, we state the following theorem:

Theorem 2 Given the open loop process

$$G(s) = \frac{K}{s^2 + a_1 * s + a_2 *} \tag{30}$$

with parameter perturbation

$$\begin{aligned} a_1 * &= a_1 - \Delta a_1 \\ a_2 * &= a_2 + \Delta a_2 \end{aligned} \tag{31}$$

If we apply the GMVCDPA, then the coordinate (a_1*,a_2*) is inside of the stability zone defined by the margins

$$\begin{aligned} \partial a_1 &= a_1 * - t_1 \\ \partial a_2 &= a_2 * - t_2 \end{aligned} \tag{32}$$

whenever (t_1, t_2) > 0, being t_1 and t_2 the coefficients of the T polynomial.

In Figure 6, we present the case shown in Figure 5, but now we use the GMVCDPA. As we can observe in Figure 6, the control system overcomes the parametric perturbation applied in the 25th second satisfactorily. The parameter perturbation is overcome appropriately due to the capacity of the GMVCDPA to update the cost polynomials on line, when it happens, also, can be observed a softer starting of the control signal. In Figure 7(a) the estimated parameters can be observed and in Figure 7(b) the cost polynomials can also be observed.

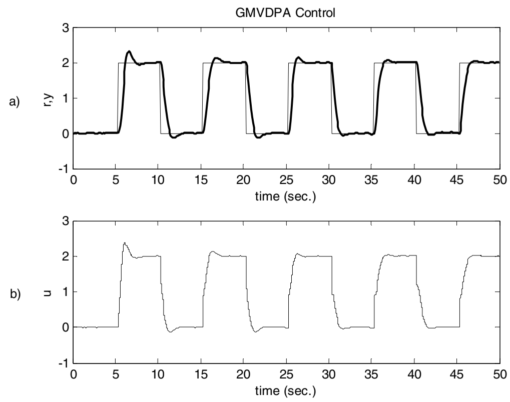


Figure 6. GMVCDPA when there is a parameter perturbation

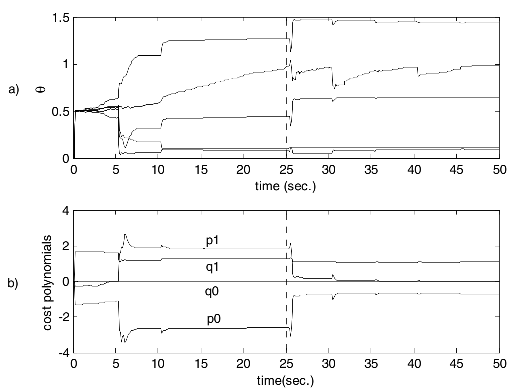


Figure 7. Parameters update in the GMVCDPA

V. IMPROVING PERFORMANCE

Another attribute of the GMVCDPA is its ability to preserve the dynamic performance originally contemplated, even in presence of perturbations. Suppose now that there is a perturbation in the parameters in closed loop,

$$PB - QA* \tag{33}$$

where A* it is the polynomial of the perturbed parameters. Considering the original design, the closed loop when the perturbation exists is

$$s^2 + t_1 s + (sa_1 * - sa_1) + t_2 + (a_2 * - a_2) \tag{34}$$

when GMVCDPA is used, according to (22), the results of differences in (29) are 0, consequently the pole assignment is completed

$$PB - QA^* = T \tag{35}$$

In order to illustrate the characteristics of the GMVCDPA in performance, we implemented practically the example 1 building the RLC circuit, and for the resistive element, we use a potentiometer to cause a change from 57Ω to 251Ω, in the 32th second. Is possible to see in Figure 8(a) the GMVCPA achieve the control in presence of the perturbation, because it tries to reduce the process output variance, but the performance is seriously affected. On the other hand, we can see in Figure 9(a), that the performance is regular, in spite of the perturbation when the GMVCDPA is used, because the action of the controller is modified to achieve the desired output [see Fig. 9(b)].

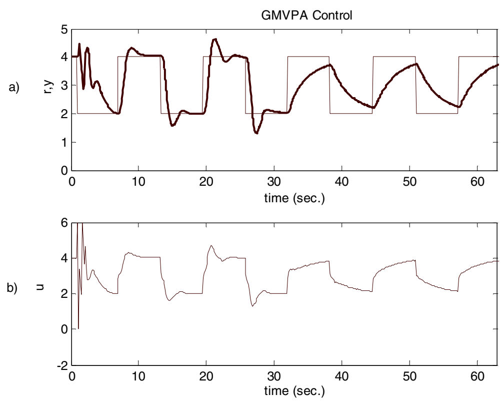


Figure 8. GMVCPA performance when there is a parameter perturbation

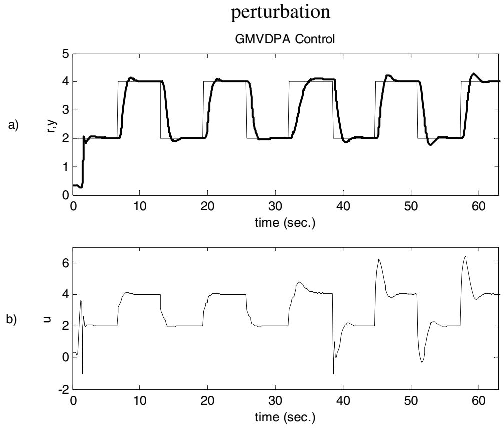


Figure 9. GMVCDPA performance when there is a parameter perturbation

VI. INDUSTRIAL APPLICATIONS

In this section, three typical industrial applications are presented, using GMVCDPA. First, for DC motor, speed and position control is implemented, and second, for a chamber prototype, the results of a temperature control are compared with those obtained with the auto-tune mode of an industrial PID controller

Speed control: in this experiment, we use a Minertia motor T03 manufactured by **Yaskawa**, which is usually used in computer peripherals as:

- **capstan and reel for magnetic tape**
- **paper and ribbon feeding for line printer**
- **X-Y plotter**
- **and small general purpose machinery**

The aim of the experiment is to control the speed of the motor. The process output is the speed variable, measured in revolutions per minute (rpm). The reference was developed in such a way that it involves ramps to be able to observe the tracking and the regulation in the same experiment. In Figure 10, we can observe that the speed of the motor follows the reference tightly.

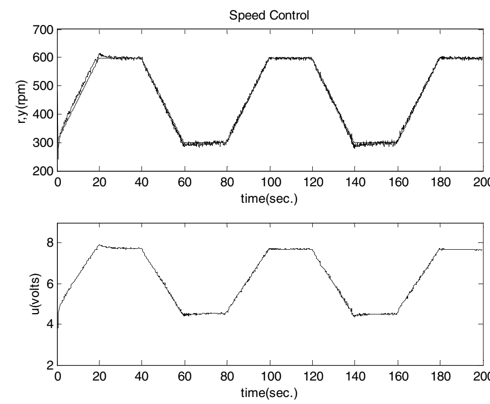


Figure 10. Motor speed control using GMVCDPA

Position control: for this experiment, we choose a small conveyor belt moved by a DC PM Gearmotor 1.61.046 manufactured by **Buehler**. The process input is the electric feeding of the motor. The process output is the belt displacement of the belt in centimeters (cm.). We use the reference shape used in the speed example. We can see in Figure 11(a), that we have remarkable tracking. Usually for position control we use a P+D Control; when the system has attained the reference value, the control signal becomes null. In this practical case, the motor has a dead zone between the -3 and 3 volts in the input process. The GMVCDPA perceives this situation, and allocates the control signal under the frontier of the dead zone. We can see in Figure 11(b) that the control signal is barely lower than the dead zone frontier, which allows a faster response before a sudden reference change.

Temperature control: In this experiment, we use a prototype which consists of in two chambers. In the first chamber there is a heating source and in the second one, there is a temperature sensor (K thermocouple). Between the chambers there is a fan that forces the air to flow between chambers (Figure 12). The process input voltage is applied to the fan.

Fan speed increase causes the temperature to increase in the second chamber. The process output is the temperature in the second chamber (°C). The process input has a dead zone between the 0 volts and 3 volts.

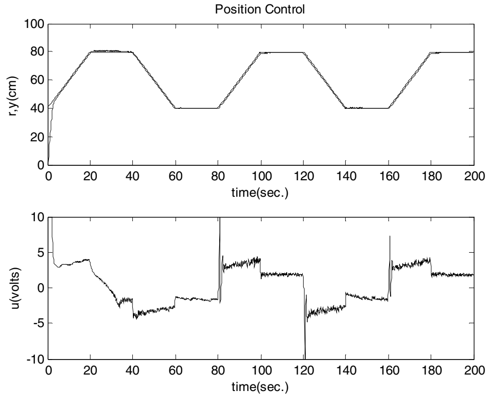


Figure 11. Conveyor belt position control using GMVCDPA

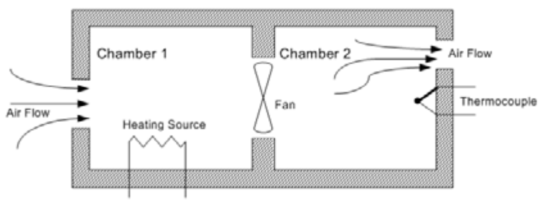


Figure 12. Temperature process

To control the process showed in Figure 12, we first use an industrial field controller, the DC1020 PID auto-tuning manufactured by **Honeywell**. When we connect the controller to the process, we select the **auto-tune** mode (ATV). The control results are showed in Figure 13(a), the reference deviation is approximately +/- 1.3°C, which for a general industrial application is considered a good performance. We can see that the control signal goes near to saturation periodically; one of the causes of such behaviour is the dead zone of the actuator and the control structure of the commercial controller.

In Figure 14 (a), we show the ability of the system to control the temperature control using the GMVCDPA. Using this self tuning controller, we improve the performance in the application. The reference deviation for this case is +/-0.3°C. In Figure 14 (b) we can see that the signal control is barely over the dead zone, allowing a better regulation.

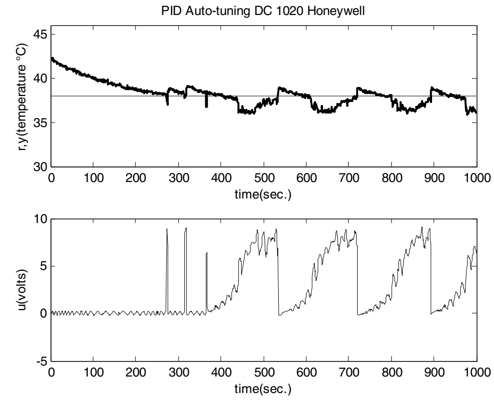


Figure 13. Temperature control using a PID Auto-tuning DC 1020 Honeywell

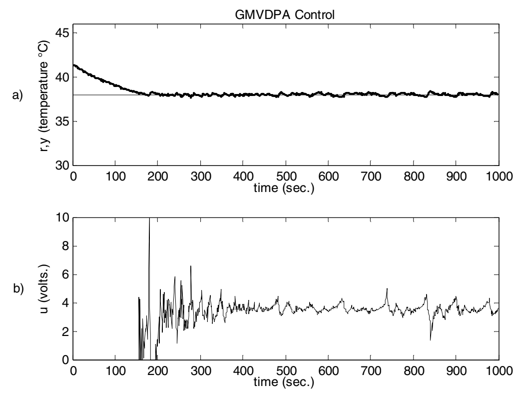


Figure 14. Temperature control using GMVCDPA

VII. CONCLUSIONS

The behavior of the GMVC and GMVCDPA which is a modified version of GMVC, was analyzed under stability and system parameter variation viewpoints. GMVCDPA provides better results than GMVC in these two areas. The laboratory experiments that were carried out with GMVCDPA, showed the outstanding performance of this controller.

VIII. REFERENCES

[l] Clarke, D.W., Gawthrop B.A. "Self-tuning controller", *IEE Proc. Control Theory Appl.*, 122, 1975, pp. 929-934.

[2] Wellstead, P.E., Zarrop M.B. *Self-tuning systems*. John Wiley & Sons, England 1991.

[3] Paz Ramos M.A., Quintero-Mármol M.E., Fernández del Busto R., "Generalized Minimum Variance with Pole Assignment Controller Modified for Practical Applications", Proceedings of the IEEE International Conference on Control Applications, Taipei, 2004, pp. 1347-1352.

[4] Poulin E., Pomerleau A., Desbiens A., Hodouin D., "Development and Evaluation of an Auto-tuning and Adaptive PID Controller", Automatica, Vol. 32, No. 1, 1986, pp. 71-82.

Automated Detection of Sunspots and Sunspot Groups in Full-Disk Solar Images

Sarab F. Majed
Department Of Mathematic, University Of Western Ontario, ON, Canada
Mehmmood A. Abd
Department Of Electrical And Computer Engineering, University Of Windsor, ON, Canada
V. Zharkova
Department Of Computing And Mathematics, University Of Bradford, BD7 1DP,UK

Abstract—**In this paper, a new adopted unsupervised segmentation technique is presented. This technique extracts both sunspots and sunspot groups from the solar disk as whole group that eases the automated sunspot group classification. MacIntosh Classification is used as standard to determine which group belongs to which class. Sunspot data are extracted from daily white light (WL) images of a solar disk captured by the SOHO/MDI satellite. New unsupervised segmentation algorithm is applied to extract sunspot groups from the full disk image with use of mathematical morphology in order to improve the quality and segregate the sunspot groups. The results were compared with the standard data from EGSO, SFC catalogues and with Mt.Wilson Report. The obtained results reveal a good accuracy of group classification, which is very promising in comparison with the manual classification, catalogues.**

***Index Terms*—Solar Disk, sunspots, edge detection, segmentation, Mathematical morphology**

1. INTRODUCTION

Sunspots observations and classification are the very essential tasks for solar astronomers. Sunspots and sunspot groups give clues to the timing of solar flares and solar activities. Sunspots are dark area appear on the photosphere and features of active regions of the sun. In addition, sunspots are born as tiny pores and some of them become developed into large spots while the other disappear [1]. Generally, sunspots have different characteristics such as size, longitudinal and shape and they are unified into groups with a large and long-lived leader spot, which has a leading polarity for the given hemisphere. The leader spots are matched with one or more trailing, smaller, spots of the opposite magnetic polarity. Such groups can be easily observed with solar magnetograph and can be realized that the ends of sunspots groups have opposite magnetic polarity [2] [3]. However, the irregularities of a sunspots shape and variability of these shapes make automated detection very difficult. The impact of the solar activity on the terrestrial atmosphere and human life is associated with a number of sunspots and the size or shape of sunspot groups. The number of sunspots and sunspot groups, Wolf number, is an indictor of the sun's magnetic field eruption. The most important indices are those called international sunspot number which are summarized by the Sunspots Index Data Center (SIDC)[4]. The precise detection of sunspot groups is a fundamental task for understanding the solar activities and their impact on The Earth's geomagnetic field and ionosphere parameters [5]. Segmentation aims to divide the images into meaningful regions that corresponding to structural units in the scene of distinguished objects of interest. Therefore, it is very important to isolate the basic features of sunspots and sunspot groups, in order to study and understand the results of the solar activity numbers. The paper is organized in the following sections. Preprocessing is presented in section 2 that includes Solar Disk acquisition and edge detection. Section 3 presents Binarization and section 4 presents sunspot groups extraction. Section 5 presents an adopted segmentation algorithm for sunspots and sunspot groups and their numbers. Finally, Section 6 presents the obtained results compared with SIDC, EGSO SFC and Mt.Wilson [3][4][6].

2. PREPROCESSING

2.1 Solar Disk Acquisition

The white and black images of a solar disk taken in Flexible Image Transport System (FITS) file format are considered in this study. These images are full solar disk calibrated synoptic daily continuum and Line of Sight (LOS) magnetogram observations. The images are taken from the Solar Oscillations Investigation Michelson Doppler Imager (SOI/MDI) aboard of the Solar Heliospheric Observatory (SOHO)[14]. The image (Figure 1) has dimensions of 1024 x 1024 pixels with 16 different intensities. The resolution is 2.6 arcsec/pixels. As in the first step, the perimeter and the solar center are detected. Then the whole solar disk is extracted from the background. Moreover, the coordinates of the centroid of a solar disk is read from the header of fits image (http://fits.gsfc.nasa.gov/) and checked by placing the solar disk diameters in perpendicular directions.

2.2 Edge Detection

Accurate solar and sunspots edge detection is important and initial task to determine to sunspots position. The position of sunspots and sunspot groups are expressed in spherical coordinates. However, diameter of the sun rely on the correct detection /calculation of solar limb. For achieving this task the threshold intensity level is found using trial and error.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_52, © Springer Science+Business Media B.V. 2010

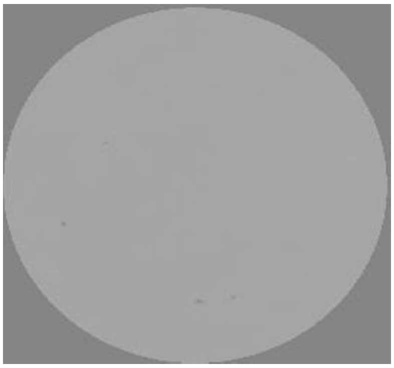


Fig. 1. The full disk image obtained from SOHO/MDI on 1 January 2000.

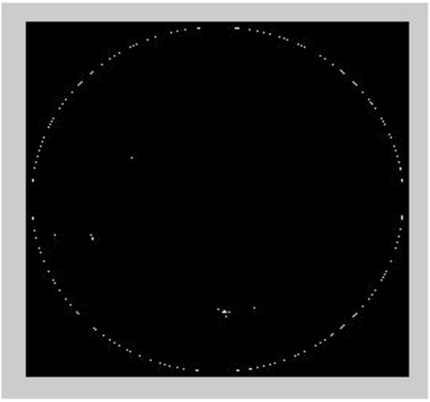


Fig. 2. Solar Edge Detection using Canny edge Detector.

Many researchers have been applied different techniques Zharkova et al (2003)[10] employed canny edge detection followed by limb fitting. Curto et al (2008) [9] were used erosion, gradient transformation and threshold. In this paper adopted Canny edge detection is used to detect the edges of solar disk. A pixel's location is declared as the edge if the value of the threshold, T, is less than the value of a gradient, g, or when g $\leq T_H$,H-high, and less than T_L, L-low. Thus, gradient components of X and Y, g_x and, g_y are computed. The edge will have the higher pixels intensity than other of surrounding regions [3]. However, after defining solar disk perimeter's pixels, the centroid needs to be found out by comparing the information from the image header with those detected from the two perpendicular maximal diameters. Consequently, all the pixels that are located outside of the solar disk are neglected.

3. NOISE REMOVING AND FEATURE ENHANCING

After defining perimeter and extracting the solar disk form the background the noise and peppers are removed. All pixels, which are outside the solar disk, are converted into black and those, which are inside solar disk, are remaining as they are. In this approach intensity filtering is employed to reduce the noise level in solar disk. Global threshold value is computed and then multiplied by *N,* where *N* is a constant varying depending on the image's contrast. Otsu's method is used to automatically perform histogram shape-based image threshold [7]. It finds the threshold that minimizes the intra-class variance, is defined as a weighted sum of variances of two classes:

$$\sigma_w^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t) \tag{1}$$

Otsu proves that minimizing the intra-class variance is as similar as maximizing inter-class variance:

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = \omega_1(t)\omega_2(t)[\mu_1(t) - \mu_2(t)]^2 \tag{2}$$

that is expressed in terms of the class probabilities ω_i, separated by threshold ,t, and the variance σ_i^2. Where classes' means, μ_i, in turn, can be updated iteratively [7]. After choosing the right threshold value, the following adopted algorithm is applied to strengthen the intensities of sunspot's penumbra and umbra and isolate them from background:

1. Read the image f (x, y)
2. While(i and j $\leq$ image's dimensions(max_x ,max_y))do
   If (*f (x, y)* >= T)then *f (x, y)*= 0;
   If (*f(x, y)* <=T)then *f(x,y)*=1 ;
   End if.

3.End

However, the output image extracted with the above algorithm is shown in Figure 3. Then after, the resulted image is converted into a binary image to find out the regions of interest, which encompasses each sunspot groups. The length of a longitudinal location of the group is used to determine which sunspot belonged to which group.

3. BINARIZATION

Binarization is the conversion from gray-scale to a binary image .One obvious way to extract the objects from background is by thresholding. Thus, in this phase, the resulted cleaned image is converted into binary image based on the obtained threshold. The threshold value is calculated using trial and error due to the variation of intensity in each sunspot. After that, the complement of the binary image is taken. In that image all the background pixels are masked to be black and all the objects keep their pixels intensities. In this case the white sunspot regions remain have wide range of pixels intensities .In other words not completely white but differ from the background. This technique is used to reserve the relations among the leading and trailers spots in each group. However, every object outside a solar disk circumference is disregarded. Figure 4 illustrates the produced resulting mask image. Consequently, all the objects attributes can be computed by using the mathematical morphology.

Morphological processing facilitates the sunspots region attributes extracting.

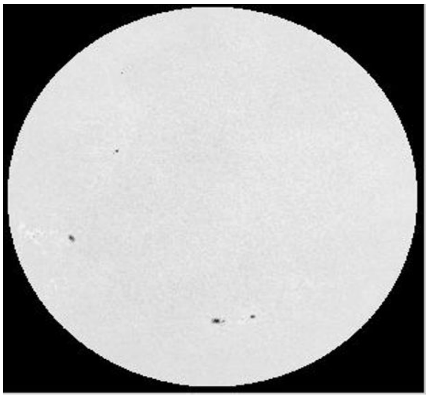


Fig. 3. Smoothed image.

Meaningful shape information of geometric objects, measuring typical characteristics area, bounding box, centroid coordinate, and the length of a perimeter. Hence, finding the sunspots coordinates (x, y), area in pixels, eases a determination of the sunspots location in relation to the solar center.

Fig. 4.The masked solar image with 23 potential sunspots.

Moreover, area of any sunspots is utilized to estimate the area in heliographic coordinates.

4. SUNSPOT GROUPS DETECTION

Extracting sunspots and organizing them into groups phase is consists of two phases: The first phase employs *Morphological operations* to enhance the image, and the second phase *gathers all related sunspots into group* based on neighborhood features of sunspots .The longitudinal of the group is used to determine which sunspot belonged to which group. The MacIntosh classification catalogue is used as standard classification in this paper [11]. The Morphological operations are used to compensate the poor quality of the images and reduce the noise similar to Zharkov *et al.*, 2005a,b[8],[12],[13]. Besides that they are used to enhance the objects /sunspots attributes. The fundamental of Morphological image processing are dilation and erosion. Dilation is an operation that utilized to thicken or grow the objects in binary images. On the other hand, the erosion shrinks the objects. The amount and the way that the objects grow or shrink is controlled by chosen selection of the structure element, which can be a set of 4 or 8 connected pixels. The structuring element that chosen in this research is an 8- connected set. However, the dilation is mathematically expressed as [3].

$$A \oplus B = \left\{ z \mid \left(\hat{B} \right)_z \cap A \neq \phi \right\} \tag{3}$$

Where A is the empty and B is the structuring element. The mathematically definition of erosion is.

$$A \Theta B = \{ z | (B)_z \cap A^c \neq \phi \} \tag{4}$$

The combination of dilation and erosion is yielding closing and opening that are higher order morphological operations. The opening is erosion of A by B, followed by dilation. Opening removes small regions and thins the filaments of object pixels. It can be expressed as.

$$A \circ B = (A \Theta B) \oplus B \tag{5}$$

While the closing morphological is dilation followed by erosion. Closing removes regions and thins filaments of background pixels. The mathematical expression of closing is expressed as.

$$A \bullet B = (A \oplus B) \Theta B \tag{6}$$

Applying the mathematical morphology operations is useful in representation and description of a region shape such as boundaries, skeletons, and convex hull. Closing fill the small gaps in a penumbra shadow to expose the umbra borders. Opening operation is utilized to opens small gaps between the touching objects. Thus, morphological operations are being useful in breaking narrow isthmuses and eliminating small (spurious) objects. The sunspots separation makes counting their number much easy [3].

Moreover, binary mathematical morphology is carried out in this study. The area, centroid, and boundaries are extracted. Then, bounding box coordinates are calculated for each spot and later used to segregate each group as independent object. Moreover, the number of objects is utilized to find out the number of sunspots in the solar disk. This number is originated from the labeling-connected components of the binary image. All region area more than five pixels consider as a spot otherwise not. The experimental executed on data set taken by SOHO/MDI. In the image which is taken on 24 January 2000 at 20: 00: UT, 59 sunspots are detected. This number is very close to maximum number that detected by Mt.Wilson [1]. However, Figure 1 shows five sunspots groups clearly detected in an image taken by SOHO/MDI on 1 January 2000. Table 1 shows a comparison between detected sunspots number in our study and EGSO (Zharkova et al., 2005)[6],[13].The false detection rate (FAR) represents the sunspots that are detected in our study but not by EGSO. The False rejection rate represents the sunspots that detected by EGSO by not detected in our study. However, the variation between the results are because image quality. However, the detection of sunspots number in this study around 80 percent, in comparison with EGSO, which could be promising results. Finally, all the adjacent sunspots that are governed by a certain number, which is varying between 3 to 15 heliographic degrees. The length of longitudinal, as that mentioned in MacIntosh Classification, is used to group sunspots together into a one group. Before computing the length of longitudinal the sunspots group must be determined as Unipolar or bipolar. Based on the conditions: " Unipolar group: A single spot, or a compact cluster of spots with the largest greatest separation between spots $\leq 3^0$. In the case of a group with penumbra (class H) ,the nearest is measured between the center of the attendant umbra and the nearest border of penumbra surrounding the main spot "[11]. Bipolar Groups: Two or more spots forming an elongated cluster of length $\geq 3^0$. Unusually there will be space near the middle of the cluster dividing it into two distinct parts. Groups with a large principal spot must be $\geq 5^0$ in length (i.e 2.5^0 plus 3^0) to be classified as bipolar " [11]. MacIntosh classification is used as standard to determine the greatest separation between sunspots in each group: " A Unipolar group with no penumbra. B - Bipolar group without penumbra on any spots.

TABLE 1
THE COMPARISON OF DETECTED NUMBERS OF SUNSPOTS IN THIS STUDY AND EGSO [13].

Date	No# Spots Our study	No#Spots EGSO	FAR	FRR
1-July-2002	23	23	0	0
2-July-2002	22	22	0	0
3-July-2002	39	41	0	2
4-July-2002	47	53	0	6
5-July-2002	38	39	0	1
6-July-2002	37	37	0	0
7-July-2002	33	32	1	0
8-July-2002	25	26	0	1
9-July-2002	17	17	0	0
10-July-2002	27	29	0	2
11-July-2002	28	28	0	0
12-July-2002	30	31	0	1
13-July-2002	53	53	0	0
14-July-2002	66	66	0	0
15-July-2002	43	43	0	0
16-July-2002	44	45	0	1
17-July-2002	58	61	0	3
18-July-2002	57	56	1	0
19-July-2002	41	40	1	0
20-July-2002	41	40	1	0
21-July-2002	41	41	0	0
22-July-2002	37	37	0	0
23-July-2002	42	42	0	0
24-July-2002	70	71	0	1
25-July-2002	86	86	0	0
26-July-2002	114	114	0	0
27-July-2002	133	138	0	5
28-July-2002	141	142	0	1
29-July-2002	118	118	0	0
30-July-2002	100	101	0	1
31-July-2002	83	83	0	0

C-Bipolar group with penumbra on one end of the group, in most cases surrounding the largest of the leader umbrae.

D–Bipolar group with penumbra on spots at both ends of the group, and with length $\leq 10^0$.
E-Bipolar group with penumbra on spots at both ends of group, and with length defined as: $10^0 \leq Lenght \leq 15^0$.
F-Bipolar group with penumbra on spots at both ends of the group, end length $\geq 15^0$. H-Unipolar group with penumbra. The principal spot is usually the leader spot remaining from a pre-existing bipolar group"[11]. However, the above-mentioned MacIntosh classification conditions are used as standard in this segmentation approach.

5. SUNSPOT GROUPS EXTRACTION ALGORITHM

Applying morphological operations enhances the sunspots characteristics and simplifies the way for determining sunspots boundaries [3]. In this paper, 8-connected components are utilized to separate each sunspot and its splinters in the single area, or group. After that, a leading sunspot is taken as the central spot from which the boundary of sunspots groups is defined. The leading spot is defined by it's area and location. The ending spot is usually larger in size and has a stronger magnetic field and located slightly closer to the equator than the trailing spots. Moreover, it is usually located at the western direction and the trailers /followers in the eastern direction. These attributes are marked with respect to the solar center. Consequently, the sub-region of sunspots is grouped in a single larger region, or as a sunspot group. Sunspots are grouped by using the length of longitudinal that confines the group as that described in MacIntosh classification. A longitude length variation for a sunspot group is ranging in $10^0 \geq legth \geq 15$ heliographic degrees. In this paper, the algorithm starts by computing the bounding box for each sunspot. The largest bounding box for a leading sunspot will be adjusted to encompass all the neighboring ones. The algorithm is iterative and terminates when no more expansion occurs or the length exceeds the maximum longitude length.

If the initial bounding box has vector $V_0 = (x_0, y_{0,} xw_0, yw_0)$ and the neighbor box has vector $V_r = (x_r, y_r, xw_r, yw_r)$ then the expanded bounding box has a vector $V = (x, y, xw, yw)$.

Where $x = \min(x_0, x_r)$ and $y = \min(y_0, y_r)$

$$xw = \max(x_r + xw_r - x, x_r + xw_r - x)$$

$$yw = (y_r + yw_r - y, y_r + yw_r - y)$$

This algorithm has to be repeated until no more bounding box expansion occurred. The output of this algorithm will be an independent image of each group. However, each group image is extracted with the detail of the original image, such as a date and time. Figure 4 shows four sunspot groups are detected, in image that taken by SOHO/MDI on 12 July 2002,in the smoothed image in Figure 3 corresponding to five of Mt Wilson. However, this difference is occurred due to the image quality and the size and location of some sunspots groups, which are very small and close to the edge of the sun. Besides, that the automated systems is based on extracting all available information from single image each day, whereas sunspots number changes by the time. Table 2 shows the accuracy of our automated sunspot groups detection and extraction in comparison with Mt Wilson for July 2002.

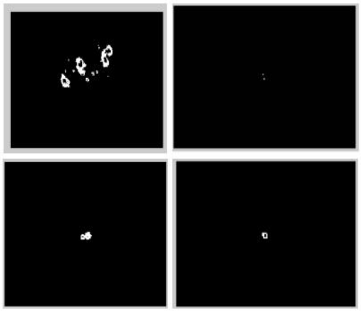


Fig.4. Four sunspot group detected in an image that taken by SOHO/MDI on 12 July 2000.

TABLE 2
THE NUMBER Of SUNSPOT GROUPS THAT ARE DETECTED FOR JULY 2002 IN COMPARISON WITH THOSE DETECTED BY MT. WILSON [10].

Date	#Spot Group Mt Wilson	#Spot group, Detected	FAR	FRR
01-July-02	5	5	0	0
02- July -02	7	7	0	0
03- July -02	10	9	0	1
04- July -02	8	10	2	0
05- July -02	11	11	0	0
06- July -02	11	10	0	1
07- July -02	6	6	0	0
08- July -02	7	7	0	0
09- July -02	9	9	0	0
10- July -02	7	7	0	0
11- July -02	-	-		
12- July -02	5	4	0	1
13- July -02	7	7	0	
14- July -02	6	7	1	0
15- July -02	9	9	0	0
16- July -02	9	8	0	1
17- July -02	7	6	0	1
18- July -02	4	5	1	0
19- July -02	5	7	2	0
20- July -02	7	7	0	0
21- July -02	8	8	0	0
22- July -02	9	8	0	1
23- July -02	10	9	1	0
24- July -02	11	10	1	0
25- July -02	9	9	0	0
26- July -02	12	10	2	0
27- July -02	15	11	4	0
28- July -02	13	11	2	0
29- July -02	10	8	2	0
30- July -02	11	9	2	0
31-July-02	8	8	0	0

Table 1 shows a comparison between our method and Mt Wilson reports results. Moreover, it is worthy to compute the false acceptance rate (FAR) (where we detect the sunspots and Mt Wilson did not detect them) and the false rejection rate (FRR) (Mt. Wilson detected the sunspots and we did not detect them). However, this variance of detection results due to the some external factors such image quality, clouds or poor visibility. From table 2 it is clear the result's accuracy of detected sunspots groups in this study around 84 percent.

6. CONCLUSION

This unsupervised segmentation technique has shown promising results that produce reliable sunspots and sunspot groups detection and extraction.

The used data set is fits format images obtained from SOHO/MDI instrument as a white light full disc image. The proposed technique shows promising segmentation and recognition results. Automotive recognition is minimizing labors needs and time consuming in comparison with manual methods.

This algorithm can be applied on any type of image format with minor modification and can be applied in biological applications. Further more, the rustles can be facilitating the automotive sunspot group detection and classification.

REFERENCES

[1] David H. Hathaway, Robert M. Wilson And Edwin J. Reichmann, " Group Sunspot Numbers: Sunspot Cycle Characteristics," Solar Physics Vol., 211,2002, pp. 357–370.

[2] Guseinov, M. D," The Polarity Of The Magnetic Field In Bright Umbral Points In Sunspots," Krymskaia Astrofizicheskaia Observatories, Izvestiia, Vol. 68, 1983, pp. 36-39.

[3] Gonzalez, R. C, Wood, R. E,"Digital Image Processing," Prentice Hall, 2004 , pp. 338-446.

[4] Observatories Royal de Belgium (2009-9-1) Sunspots Index Data Center (SIDC) [Online]. Available: http://www.icsu-fags.org/ps11sidc.htm

[5] Mitic, M. Djordjevic ,A. Mihajlovic,S. J. "Changes In atmosphere and Ionosphere Caused By Solar-geomagnetic Activity ,Within Belgrade Area ," ," EGS XXVII General Assembly, Nice,(21-4-2009) [Online]. Available: http://adsabs.harvard.edu/abs/2002EGSGA..27..631M

[6] Zharkova et al., Solar Physics.(1-5-2009). European Grid of Solar Observation, Solar Feature Catalogues [Online]. Available: http://solar.inf.brad.ac.uk/pages/ss_start.php

[7] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," Journal of Electronic Imaging. Vol.13 (1) ,2004, pp. 146–165

[8]. S. Zharkov, V. Zharkova, S. Ipson, A. Benkhalil," The Technique For Automated Recognition On Full Disk Solar ," EURASIP Journal .Vol. 15, 2005, pp. 2573-2584

[9] J. J. Curto, M. Blanca and E. Martínez, "Automatic Sunspots Detection on Full-Disk Solar Images using Mathematical Morphology," Solar Physic. Vol. 250, 2008, pp. 411-429.

[10] Mount Wilson Observatory, (13-12-2008) [Online]. Available: http://www.mtwilson.edu/hk/ARGD/Sunspot_Cycle/

[11]. Patrick S. MacIntosh, (1990)," The Classification Of Sunspot Groups ", Solar Physics Vol. 125,1990, pp. 251-267.

[12]. Zharkov, S., Zharkova and V.,Ipson, S., 2005b," Statistical properties of sunspots in 1996-2004 : I. Detection ,North-South asymmetry and area distribution ," Sol. Phys. ,228, PP. 377-397.

[13]. Zharkova, V., Aboudarham, J., Zharkov, S., Ipson, S., Benkhalil, A., Fuller, N., " Solar feature Catalogue in EGSO," Sol. Phys., Vol. 228, 2005, PP 361-375.

[14]. Solar Heliospheric Observatory (SOHO), (10-9-2009, [Online]. Available:http://sohowww.nascom.nasa.gov/

Stochastic Model Based Approach for Biometric Identification

Mofakharul Islam[1] Sitalakshmi Venkataraman[2] Mamoun Alazab[3]
Internet Commerce and Security Laboratory
Graduate School of Information Technology and Mathematical Sciences
University of Ballarat, Australia
mo.islam@ballarat.edu.au[1]
s.venkatraman@ballarat.edu.au[2]
m.alazab@ballarat.edu.au[3]

***Abstract*- In this paper, we present a new stochastic model based approach for enhanced image segmentation in biometric identification systems. Biometric features such as fingerprint, face, iris, hand geometry and more recently dental features are being used for human identification. Image analysis of each of these biometric features has various challenges to overcome. To address such contemporary problems of image segmentation, we provide a novel approach based on *maximum a posteriori* (MAP) fitting Gaussian mixture model using Expectation-Minimization (EM) algorithm within the Bayesian framework. Our new algorithm captures the pixel intensity by the likelihood term in Bayesian Networks, and *a priori* biasing term of the spatial location information with the help of Markov Random Fields (MRF) model. We have employed a novel approach of using Daubechies wavelet transform for texture feature extraction that uses MRF model and a robust technique of determining the number of pixel classes based on Cluster Ensembles for a reliable segmentation of dental X-ray images. We present how our approach could be applied in dental biometrics to achieve very fast and reliable human identification. Experiments show that our new unsupervised image segmentation technique provides accurate feature extraction and teeth segmentation for effective biometric identification.**

I. INTRODUCTION

Human physical characteristics such as fingerprint, face, retina scans, iris patterns, hand geometry, and others have been used in various personal identification systems [1], [2]. However, in order to meet the specific requirements of different biometric applications, more recently, researchers have been probing into other internal features such as teeth, spine and chest contours [3], [4]. Since internal features such as iris and retina consist of soft tissues that could get damaged or deformed, recent research studies have established that dental features form a more reliable candidate for biometric identification as they are more robust and could be preserved for a sufficiently long period of time, even after death [5], [6]. Another motivation is that similar to medical diagnosis, X-ray image processing approaches could be adopted as low-cost solutions for personal identification systems using human internal features such as teeth. Hence, in this research, we concentrate on the image segmentation approaches suitable for dental biometric identification systems that are useful not only for ante mortem human identification but also for post-mortem identification, due to their survivability characteristic for a long period of time.

In biometric identification, where dental characteristics such as spacing between teeth, teeth shapes, sinus patterns and root / crown morphologies are used to identify human individuals, image segmentation and classification is a complex task due to the non-standard regions of interest (ROI) and information present in dental X-ray images. The content variations are also attributed to the quality of contrast, non-uniform intensities and digitisation / high frequency noise that are inherent in X-ray images. Hence, traditional image segmentation algorithms such as edge-based, region-growing and threshold techniques that have been successful in certain applications are not suitable for X-ray images [7], [8]. Similarly, contemporary approaches, which combine region growing methods with edge-detection techniques or even statistical model based approaches, are either computationally expensive or more suitable for colour models [9], [10].

In order to address the above mentioned problems, this paper proposes a novel stochastic model based approach to perform image segmentation effectively. The application of our proposed approach in biometric identification using dental images has been experimentally tested to produce effective and accurate results for feature extraction and image segmentation. The rest of the paper is organised as follows. In Section II , we present a modified classical finite mixture model that employs Markov Random Fields (MRF) model with Expectation-Minimization (EM) algorithm under Bayesian framework. Section III describes our novel approach of image segmentation of dental X-ray images where the gray-level intensity and texture features play crucial role in pixel clustering. Texture feature extraction are performed using Daubechies wavelet transform and ensemble of clustering to arrive at reliable number of components/segments in an image under unsupervised framework. In Section IV, we discuss the experimental results of our proposed approach applied on teeth X-ray images and in Section V we present our conclusions and directions for future work.

II. MAP FITTING GAUSSIAN MIXTURE MODEL

Among the stochastic model based methods, Gaussian mixture model or the finite mixture model has received significant attention from researchers in the last decade due to its improved segmentation results when compared to other image segmentation techniques [11], [12]. A common

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_53,

drawback of such statistical model based methods is their high computational intensity. However, adapting a finite mixture model for image segmentation is recently becoming popular [13]. We propose a stochastic model based on classical finite mixture model that employs Markov Random Fields (MRF) to capture the neighbourhood relationships among the neighbouring pixels. The proposed model has some advantages, especially in the case of contextually dependent physical phenomena like images, as it incorporates the outstanding features of the MRF model without additional computational burden. In the proposed method, MRF introduces *a priori* distribution that takes into account the neighbourhood dependency or relationship among the neighbouring pixels. We elaborate a stochastic model based segmentation framework based on Gaussian mixture models below.

The proposed model is an introduction of MRF model into the classical mixture model using Expectation-Minimization (EM) algorithm under Bayesian framework. Recently, approaches using MRF models and Bayesian methods have provided answers to various contemporary problems in image segmentation. The theoretical framework relies on Bayesian estimation via combinatorial optimization (Simulated Annealing). The application of a clustering method to image segmentation has the particular characteristics that spatial information should be taken into account. Here, in addition to intensity-values of pixels, the pixel location must also be used to determine the cluster to which each pixel is assigned. Conceptually, in most cases, it is desired that the same cluster label be assigned to spatially adjacent pixels. In implementing these ideas, the Bayesian framework provides a natural approach to address the problem of image segmentation. Pixel intensity information is captured by the likelihood term in Bayesian Networks, while *a priori* biasing term captures the spatial location information with the help of MRF model.

In dental biometrics, where dental features are required to be extracted from X-ray teeth images, the existing image segmentation algorithms are ineffective if they are unable to handle the inherent lighting and imaging conditions of X-ray images. We propose an adaptive Gaussian mixture model that employs MRF to handle these changing conditions in X-rays. The conceptual idea behind the MRF model is that the neighbouring pixels must have influences on a pixel. What we capture through the MRF model is a pixel's interaction with its neighbouring pixels. In other words, we adapt the MRF model to capture the spatial relationship among the neighbouring pixels. As a result, it determines the probability of a pixel to remain in the same cluster with its adjacent pixels. In this way, the spatial information captured addresses the various X-ray imaging artefacts such as noise and non-homogeneity in intensity, as the neighbourhood information gathered is used to achieve a higher degree of discriminative ability in determining different pixel clusters or segments. We provide below our mathematical model in adapting the MRF approach through the introduction of neighbourhood relationship among the pixels in addition to the pixel features in order to efficiently handle X-ray images towards achieving increased accuracy in image segmentation.

We present here a mixture model with an unknown number of components *K*, each of which has its own vector of density parameter θ_j for each feature. Let x^i denote the observation at the *i*th pixel of an image (*i=1, 2,…..N*) modeled as independent and identically distributed (iid). So, the probabilities of the *i*th pixel belonging to the *j*th class label are given as follows:

$$\eta_j^i = P(j \mid x^i) \quad (1)$$

We assume that the density function $f(x_i \mid \Phi, \Gamma)$ at any observation, x_i is given by the mathematical expression as follows:

$$f(x_i \mid \Phi, \Gamma) = \sum_{j=1}^{K} p_j^i \psi(x_i \mid \theta^j) \quad (2)$$

where, $\psi(x_i \mid \theta^j)$ is a Gaussian distribution with parameters $\theta_j = (\mu_j, \sigma_j)$, and *K* is the number of components in the mixture.

We introduce *Maximum a posteriori* (MAP), a prior distribution for the parameter set Φ that takes into account spatial information based on the Gibbs function. According to the Hammersley-Clifford theorem [14], [15], *Gibbs distribution* takes the following form:

$$P(\Phi) = \tfrac{1}{Z} \exp(-\beta \sum V_{N_i}(\Phi)) \quad (3)$$

where, Φ is a vector of features, β is called the regularization parameter and the normalizing constant *Z* is called a partition function. $V_{N_i}(\Phi)$ denotes clique potential of the label configuration p^m within the neighbourhood N_i of the *i*th pixel which can be calculated as;

$$V_{N_i}(\Phi) = \sum_{m \in N_i} g(u_{i,m}) \quad (4)$$

where, $u_{i,m}$ denotes the distance between the two label vectors p^i and p^m . The function *g(u)* must be gradually increasing and non-negative. The neighbourhood N_i consists of horizontal and vertical neighbours of *i*.

To calculate the posterior probability distribution of one random variable while the value of another is given, we make use of Bayes theorem. According to Bayes theorem [14], the posterior probability density function (PDF) is defined as follows:

Posterior PDF = <u>*Likelihood Term x Prior Probability*</u>

Normalising Constant

Then, Posterior PDF for a random variable X, given that the data $Y = y$ takes the form as follows:

$$f_{X|Y=y}(x) = \frac{f_X(x)L_{X|Y=y}(x)}{\int_{-\infty}^{\infty} f_X(x)L_{X|Y=y}(x)\,dx}$$

where, $f_X(x)$ is the prior density of X and the likelihood function as a function of x is given by

$L_{X|Y=y}(x) = f_{Y|X=x}(y)$,

$\int_{-\infty}^{\infty} f_X(x)L_{X|Y=y}(x)\,dx$ is the normalizing constant, and $f_{X|Y=y}(x)$ is the posterior density of X, given the data $Y = y$.

Now, we can derive a posterior log density function as follows:

$$P(\Phi,\Gamma \mid X) = \sum_{i=1}^{N} \log(x^i \mid \Phi,\Gamma) + \log P(\Phi) \qquad (5)$$

The *Expectation-Maximization* (EM) algorithm requires that the computation of the conditional expectation value, z_j^i of the hidden variables must be at the Expectation Step (E-step). The E-step computes the expected value using current estimate of parameter and the observed data in the image [16]. Due to the application of the MAP estimation of the parameters $\{p_j^i\}$ and $\{\theta_j\}$, the mathematical expression for z_j^i takes the form as given below:

$$z_j^{i^{(t)}} = \frac{p_j^{i^{(t)}} \psi(x^i|\theta_j^{(t)})}{\sum_{l=1}^{K} p_l^{i^{(t)}} \psi(\chi^i|\theta_j^{(t)})} \qquad (6)$$

Next, the maximization of the following log-likelihood corresponding to the complete data set is performed in the Maximization Step (M-step) with t denoting the iteration step:

$$Q_{MAP}(\Phi,\Gamma \mid \Phi^{(t)}\Gamma^{(t)}) \qquad (7)$$

The M-step determines the maximum likelihood estimate of the parameter using the data from the E-step [16]. For each parameter, Q_{MAP} can be maximized independently, which yields the next iteration update equations for parameter of the component densities $\mu_j^{(t+1)}$ and $[\sigma_j^2]^{(t+1)}$. Thus, the problem of local minima in the EM algorithm has been successfully addressed by using *Simulated Annealing.*

III. Proposed Dental X-ray image Segmentation Approach

In different imaging applications like medical diagnosis, biometrics, security and surveillance, mineral and mining, and material science, where the primary goal is accuracy in image segmentation, making use of subtle information related to intensity and texture could play a significant role in achieving this goal. Hence, in this research work, we have effectively utilized the pixel's gray level intensities and texture features for clustering purpose. Here, Daubechies wavelet transform has been employed as texture descriptor. We combine i) the gray level intensity, ii) Daubechies transform wavelet descriptor, iii) Cluster Ensembles, a robust number of component / segment finding tool, where the SNOB [17], [10] is used as a starting point, and iv) the proposed MAP fitting Gaussian mixture model, into an integrated segmentation model. The main purpose of such a novel image segmentation approach is to achieve greater accuracy in meaningfully segmented components that is required for a wide variety of industrial images.

We adopt the following four steps for the experimental investigation of our proposed dental X-ray image segmentation model:

Step 1: Extract the texture features,
Step 2: Find the number of components,
Step 3: Determine the cluster ensembles, and
Step 4: Segment the image using the proposed MAP Fitting Gaussian Mixture Model

Step 1: Extract the Texture Features

Texture characterizes local variations of colour or intensity in an image. The neighbourhood property is a common similarity among all the texture descriptions. Statistical methods consider texture as random phenomena with identifiable local statistics. So, we can define texture as an image feature which is characterized by the spatial distribution of gray values (for X-ray images) or colour patterns (for colour images) in a neighbourhood surrounding the pixel.

As the segmentation quality greatly depends on the choice of feature (intensity, colour, texture, and co-ordinate) descriptor, in our proposed approach, we have utilized pixel's gray level intensity along with texture feature in the analysis of contextually dependent physical phenomena of X-ray images. Hence, in this research work, we adopt a more context sensitive image segmentation technique using wavelet transform to produce reliable and accurate segmentation of X-ray images based on subtle intensity and texture variation. Although, wavelet transform has made significant contribution in several areas of image processing such as image enhancement, image compression, and image registration, its use in image segmentation is very limited and is gaining attention recently [18-21]. Islam et al. [21] have used wavelet transforms in their image segmentation approach to extract texture features from natural colour images and have reported promising results. However, image analysis in dental biometrics demands a more sensitive texture descriptor that could provide an efficient way to capture the subtle variations in texture. The Daubechies wavelet transform has the capability to better represent image semantics such as intensity variations and segment configurations [22], [23]. Hence, in this research, we have used Daubechies wavelet transform as the texture descriptor under Markovian framework for the purpose of image segmentation as it is found to be more accurate and sensitive in capturing texture details in an X-ray image.

Step 2: Find the Number of Components

Existing approaches are found to be either incapable of handling data with noise or do not perform well in handling discrete data like image data [7], [10]. Our experimental study indicates that an unsupervised data clustering approach is ideal for dental X-ray images and is found to be more effective in handling noisy image data that have diverse varieties of faults and imperfections.

After the extraction of the texture features, in this step, we input all these features along with pixel's intensity value into the Snob program [17] to determine the number of classes present in the X-ray image. We have used the Snob program only as a starting point to find the initial number of components. In general, Snob has been found quite effective in identifying the number of components in an image but sometimes it produces spurious results due to its limitations to handle some imaging artefacts like noise, shadow, and intensity non-homogeneity. Hence, we refine the number of components by determining the cluster ensembles in the next step.

Step 3: Determine the Cluster Ensembles

In this step, we devise a robust unsupervised technique to find the number of components in an image automatically. An emerging approach like *Ensembles of Clusterings* could play a significant role to augment the technique for finding accurate and authenticated number of components in an image. Ensembles of clustering are basically a combination of multiple clusterings used to arrive at a consolidated clustering without accessing the original data and algorithms. Here, only symbolic cluster labels are mutually shared. Multiple clusterings can be produced either using a single clustering algorithm by varying the different parameters or using different clustering algorithms on the same data set and parameters [24], [25]. With Snob as a starting point for getting initial clusterings, we base upon Ensembles of Clusterings to produce a consolidated clustering. In our experimental procedure, we produced multiple clusterings for different parameter values and then selected the stable solution found in the plot of the number of clusters. The convergence criteria or number of iterations of MRF model is the main parameter that we have varied to get multiple clusterings.

Step 4: Segment the Image Using the Proposed MAP Fitting Gaussian Mixture Model

Next, the four original features, such as intensity / gray scale level, and texture, along with the number of classes obtained from Ensembles of Clusterings, become the input to our segmentation model as discussed in Section II. Accordingly, clustering of pixels is done on the basis of homogeneity in intensity and texture properties. As a result, regions are obtained where there is a discontinuity in either intensity or texture.

In our segmentation model, the K-means algorithm determines the centres of the clusters for each component of the image. Then the Expectation Maximization (EM) algorithm is used for estimating the model parameters. In fact, model parameter estimation, incorporating the neighbourhood relationships among pixels and segmentation of pixels are run simultaneously within MAP estimation under Markovian framework. The whole process is being done using a particular iterative procedure in an interleaved manner. In doing so, MAP classifies the pixels of the input images into different pixel classes on the basis of intensity and texture features. Here, the function Q_{MAP} plays a crucial role in determining the actual membership of a class. In fact, the intensity and texture features of pixel directly influence the maximization of the Q_{MAP} function. That means every individual feature has some influence on the maximization process of the Q_{MAP} function. Some features have greater influences while others have minimum effect. If the influence of a particular feature is greater than others in the maximization process of the Q_{MAP} function for a certain pixel, then the clustering of that pixel is greatly influenced by that particular feature.

IV. Model Significance and Experimental Results

Recognition of a particular object in an image is a task that is carried out as pixel classification and the most significant aspect in pixel classification is the feature extraction. Feature extraction is a process of measurement and it forms the main role in the process of object recognition or image segmentation. The performance of image segmentation in an experimental study depends strongly on the choice of feature extraction approach.

Our approach in image segmentation is different from previous work in a few major ways. We find in existing literature, approaches that adopt pre-processing procedures such as gum-line detection, pixel classification and adaptive thresholding are based on classical techniques like segmentation using edge, line, contour and region [26-28]. The algorithms that are proposed in other previous work are exclusively based on high level features using classical approaches that are computationally more expensive. They have not used any stochastic or statistical models that provide maximum accuracy. In our novel approach, we have developed a stochastic model to make it efficient for pixel based image segmentation. Though some research has been conducted in using stochastic model based approach for processing magnetic resonance (MR) images of human organs for medical diagnosis [29], most of these existing approaches are greatly influenced by similarity in intensity. In dental images, mere similarity in pixel's intensity does not really mean they belong to the same object. In the front of background intensity of dental images, many objects of similar intensity may lie, which can be characterized through variation of their textures.

In our proposed approach, we have used the Daubechies wavelet transform for texture feature extraction as it has the ability to capture subtle texture information with better computational efficiency. Some researchers have used directional wavelet transform and image gradient averaging techniques for enhancing X-ray images [18], [19]. However, as mentioned before, such intensity-based enhancement techniques are not suitable for dental images. Wavelet transforms are yet to be employed as a texture descriptor in teeth X-ray image segmentation scheme that uses MRF model theory and hence our approach is unique. In addition, we have introduced a robust technique for automatic determination of the number of pixel classes in images based on Cluster Ensembles, which can be considered as a novel approach in the context of finding the number of different segments in dental X-ray image.

Regarding the experimental study of our proposed approach, our primary aim is to verify that the algorithm is able to identify different meaningful segments or regions in dental X-ray images accurately. The effectiveness of the algorithm is based on whether the algorithm takes into account the subtle variations in gray-level intensity and texture and thereby able to produce segmentation in a precise manner. While finer image segmentation has a great potential in many application domains including bio-medical, bio-technical, material sciences, mining as well as oil and gas industries, in this research we concentrate on dental X-ray images as we wish to explore the use of dental image segmentation for biometric identification.

In our proposed segmentation approach, we have made use of the fact that gray-level intensity and texture based segmentation are capable of producing sharp boundaries and homogenous regions in dental images. In fact, individual objects can be detected more efficiently with the combined features of gray-level intensity and texture than when either the texture or intensity feature is used separately. As a result, our segmentation approach with combined gray-level intensity and texture outplays others in terms of sharpness and homogeneity. The number of classes has been determined initially, which has been taken as input for the proposed segmentation model where the parameter estimation and pixel clustering are done iteratively in an interleaved manner simultaneously during the segmentation process. A series of image segmentation experiments have been conducted on real-life teeth X-rays to evaluate the effectiveness of the proposed method. Segmented images as sample outputs of our model are illustrated in Fig. 1.

In Fig. 1, we present 3 segmented images using our proposed model as against their original images. We observe that our approach has identified 5, 6, and 4 distinct regions in Image I, Image II, and Image III respectively, which was difficult to be identified with existing edge-detection or region-growing methods due to the noise present in the original data. Hence, our approach has the ability to perform beyond the classical approaches of image segmentation that are unable to capture subtle variations in intensity and texture in dental X-rays, especially with noise.

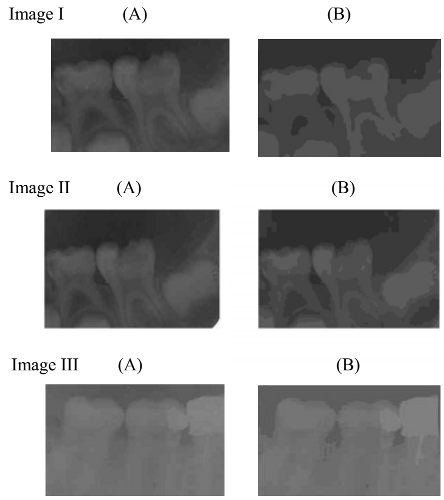


Fig. 1. (A) Original X-ray dental image and (B) Segmented dental image

Further, a quantitative performance evaluation with another similar existing method [20] has been done using a set of criteria in the form of an evaluation matrix from previous existing studies [30]. We input the segmentation results of our proposed approach as well as results from previous existing method for generating the evaluation matrix. The output from these evaluation criteria at each of the three stages is a set of scores as shown in Fig. 2. The overall scores are determined based on criteria such as segmentation accuracy and correct number of objects identified in an image. From Fig. 2, the higher scores of our proposed approach as compared to existing method clearly confirm that our proposed approach is able to segment the images more efficiently. Our experiments conducted on real-life dental X-rays reveal that the proposed algorithm is also capable of capturing the finer details of an image in terms of its gray-level intensity and texture accurately. Eventually, these regions/segments that represent internal structures of human teeth can be utilized as biometric digital data for identification purposes.

Image I performance results:

	Proposed Approach	**Existing Method**
Stage 1	0.8151	0.7305
Stage 2	0.6985	0.6147
Final Stage	0.8359	0.6914

Image II: performance results:

	Proposed Approach	**Existing Method**
Stage 1	0.8173	0.7458
Stage 2	0.7541	0.6879
Final Stage	0.69723	0.6528

Image III: performance results:

	Proposed Approach	**Existing Method**
Stage 1	0.8147	0.7639
Stage 2	0.6987	0.5825
Final Stage	0.6754	0.5769

Fig. 2. Comparative evaluation of proposed approach

V. Conclusions and Future Work

In this paper, we have proposed a novel dental image segmentation approach and have investigated its use in dental biometric identification. Our stochastic model based approach utilizes a Markov Random Field model to capture the relationships among the neighboring pixels and integrate that information into the Expectation Maximization model fitting Maximum a Posteriori algorithm. The paper has described unsupervised means of automatic discovery of classes or clusters in images to produce precise segmentation of dental X-ray images using intensity and texture information along with neighbourhood relationships among image pixels. Such an approach has provided the required accuracy in segmentation as well as in finding the number of components in an image automatically through the use of Cluster Ensembles and MRF model. In addition, our results are enhanced by the use of Daubechies wavelet transform for increasing the content sensitivity of the segmentation model, thereby addressing even noise in images. Finally, the iterations of the algorithm have been converged using specific criterion on model parameters as they were calculated for arriving at the resulting image segments of dental X-rays in an interleaved manner. A pilot experimental study has tested the algorithm on real-life X-ray images and has provided convincing results.

Our future work will test the performance of our proposed approach described in this paper using large data sets so as to determine the threshold level of noise in image that could still be catered within the model. Future work also entails in addressing the important research question with regard to the practical implementation of human dental identification in real-life biometric applications.

References

[1] J. D. Woodward, N.M Orlans and P.T Higgins, *Biometrics: Identity Assurance in the Information Age*, Berkeley, CA: McGraw-Hill/Osborne, 2003

[2] J. R. Vacca, *Biometric technologies and verification system*. Oxford: Butterworth-Heinemann, 2007.

[3] L.R. Long, S. Antania, D.J. Leeb, D.M. Krainakc, and G.R. Thomaa, "Biomedical Information from a National Collection of Spine X-rays – Film to Content-based Retrieval", *Proceedings SPIE* , 2003

[4] A. K. Jain and H. Chen, ``Matching of Dental X-ray Images for Human Identification," *Pattern Recognition*, Vol. 37, No. 7, pp. 1519-1532, 2004.

[5] G. Fahmy, D. M. Nassar, E. Haj Said, H. Chen, O. Nomir, J. Zhou, R. Howell, H. H. Ammar,M. Abdel-Mottaleb, and A. K. Jain, "Toward an Automated Dental Identification System," *Journal of Electronic Imaging*, Vol. 14, No. 4, pp. 1–13, 2005.

[6] S. Purushotham, and M. Anouncia, "Enhanced Human Identification System Using Dental Biometrics", *Proceedings of the 10th WSEAS International Conference on Neural Networks*, Prague, Czech Republic, 2009, pp. 120 -125, 2009.

[7] C.S. Wallace, and D.M. Boulton, "An Information Measure for Classification", *Computer Journal*, Vol. 11, No. 2, pp. 185-194, 1968.

[8] N. R. Pal and S. K. Pal, "A Review on Image Segmentation Techniques", *Pattern Recognition*, Vol. 26, No. 9, pp. 1277 - 1294, 1993.

[9] Shih, F.Y. and Cheng, S., "Automatic Seeded Region Growing for Color Image Segmentation", *Image and Vision Computing*, Vol. 23, No. 10, pp. 844–886, 2005.

[10] C.S. Wallace, and D..L. Dow, "MML Clustering of Multi-state, Poisson, Von Mises Circular and Gaussian Distribution", *Statistics and Computing*, Vol. 10, No. 1, pp. 73-83, 2000.

[11] G. McLachlan and D. Peel, *Finite Mixture Models.* New York: John Wiley&Sons, 2000.

[12] P. Meinicke and H. Ritter, "Resolution-Based Complexity Control for Gaussian Mixture Models," *Neural Computation,* vol. 13, no. 2, pp. 453-475, 2001.

[13] M. Alf`o, L. Nieddu and D. Vicari, "A Finite Mixture Model for Image Segmentation", *Statistics and Computing*,, Vol. 18, No. 2, pp. 137 – 150, 2008.

[14] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs distribution, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 6, pp. 721–741, 1984.

[15] P. Clifford, "Markov Random Fields in Statistics", In: Geoffrey Grimmett and Domnic Welsh (Eds.), Disorder in Physical Systems: A Volume in Honour of John M. Hammersley, pp. 19–32. Oxford University Press, 1990.

[16] T.K. Moon, "The Expectation Maximization Algorithm", *IEEE Signal Processing Magazine*, Vol. 13, No. 6, pp. 47-60 1996.

[17] D.M Boulton,. and C.S Wallace, "A Program for Numerical Classification", *The Computer Journal*, Vol. 13, No.1, pp. 63–69, 1970.

[18] D. Heric and B. Potocnik, "Image Enhancement by Using Directional Wavelet Transform", *Journal of Computing and Information Technology*- Vol. 14, No. 4, pp. 299–305 .2006.

[19] J. Feng, N. Xiong and B. Shuoben, "X-ray Image Enhancement Based on Wavelet Transform,", *Proceedings of IEEE Asia-Pacific Services Computing Conference, APSCC*, pp. 1568-1573, 2008.

[20] M. Islam, J. Yearwood and P. Vamplew "Unsupervised Color Textured Image Segmentation Using Cluster Ensemble and MRF Model" , *Advances in Computer and Information Sciences and Engineering*, Springer Netherlands, ISBN 978-1-4020-8740-0 (Print) 978-1-4020-8741-7 (Online), pp. 323-328, 2008.

[21] M. Islam, P. Vamplew and Yearwood J., "MRF Model Based Unsupervised Color Textured Image Segmentation Using Multidimensional Spatially Variant Finite Mixture Model", *International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 2008)*, Bridgeport, USA, 2008.

[22] I. Daubechies, "Orthogonal Bases for Compactly Supported Wavelets", *Communications on Pure and Applied Mathematics*, Vol. 41, pp. 773-789, 1988.

[23] J.Z. Wang, G. Wiederhold, O. Firschein and S.X. Wei, "Content-based Image Indexing and Searching Using Daubechies' Wavelets", *International Journal of Digital Library*, 1, pp. 311-328, 1997.

[24] A.L.N. Fred, and A.K. Jain, "Data Clustering Using Evidence Accumulation", *Proceedings of the 16th International Conference. on Pattern Recognition 'ICPR'02'*, Quebec, Canada, 2002.

[25] A. Strehl, and J. Ghosh, "Cluster Ensemble - A Knowledge Reuse Framework for Combining Multiple Partitions", *Journal of Machine Learning Research*, Vol. 2, pp. 583-617, 2002.

[26] H. Chen and A. K. Jain, "Tooth Contour Extraction for Matching Dental Radiographs," *Proceedings of 17th International Conference. on Pattern Recognition 'ICPR'04*,' Cambridge, UK, 2004.

[27] J. Zhou and M. Abdel-Mottaleb, ``Automatic Human Identification Based on Dental X-ray Images," *Proceedings of SPIE Technologies for Homeland Security and Law Enforcement Conference, the Biometric Technology for Human Identification Conference*, Orlando, USA, 2004.

[28] E. Haj Said, D. M. Nassar, and H. H. Ammar, "Image Segmentation for Automated Dental Identification," *SPIE Electronic Imaging*, San Jose, CA, 2006.

[29] Y. Zhang, Michel Brady and S. Smith, "Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation Maximization Algorithm," *IEEE Transactions on Medical Imaging*, Vol. 20, No. 1, pp. 45- 57, 2001.

[30] Islam, M., "Unsupervised Color Image Segmentation Using Markov Random fields Model", *Master by Research Thesis*, Graduate School of Information Technology and Mathematical Sciences, University of Ballarat, Australia, 2008.

A New Stochastic Model Based Approach for Object Identification and Segmentation in Textured Color Image

Mofakharul Islam[1] Paul A. Watters[2]
Internet Commerce and Security Laboratory
Graduate School of Information Technology and Mathematical Sciences
University of Ballarat, Australia
mo.islam@ballarat.edu.au[1]
p.watters@ballarat.edu.au[2]

***Abstract*-We investigate and propose a novel stochastic model based approach to implement a robust unsupervised color image content understanding technique that segments a color textured image into its constituent parts automatically and meaningfully. The aim of this work is to detection and identification of different objects in a color image using image segmentation. Image segments or objects are produced using precise color information, texture information and neighborhood relationships among neighboring image pixels. As a whole, in this particular work, the problem we want to investigate is to implement a robust Maximum a posteriori (MAP) based unsupervised color textured image segmentation approach using *Cluster Ensembles*, MRF model and Daubechies wavelet transform for identification and segmentation of image contents or objects. In addition, *Cluster Ensemble* has been utilized for introducing a robust technique for finding the number of components in an image automatically. The experimental results reveal that the proposed model is able to find the accurate number of objects or components in a color image and can produce more accurate and faithful segmentation of different meaningful objects from relatively complex background. Finally, we have compared our results with another similar existing segmentation approach.**

I. INTRODUCTION

Naturally, color image segmentation demands well defined borders of different objects in an image. So, there is a fundamental demand of accuracy. The segmented regions or components should not be further away from the true object than one or a few pixels. So, there is a pressing need for an improved image segmentation technique that can segment different components precisely.

Image data have some particular characteristics that differentiate them from other form of data. Image data may have corrupted values due to the usual limitations or artifacts of imaging devices. Noisy data, data sparsity, and high dimensionality of data create difficulties in image pixel clustering. As a result, image pixel clustering becomes a harder problem than other form of data. Although there are some existing algorithms for unsupervised color image segmentation, none of them has been found to be robust in determining an accurate number of components or segments and eventually in producing faithful segmentation.

More noise means more uncertainty. Handling uncertainty is the most significant problem in image segmentation. Conventional probability theory was the primary mathematical model that used to deal with this sort of uncertainty Researchers have found that probability theory alone is not quite adequate to handle uncertainty especially where the situation demands more precise handling. As a result, there was a need to augment conventional probability models by introducing some additional concept or knowledge to make it more efficient and effective. Over the last few decades researchers from the different computing communities proposed various approaches that are quite efficient in handling uncertainty. Markov Random Field Model (MRF) theory based approach is one of them. The only difference between MRF model based approaches with other models is that MRF model based approaches consider the neighborhood influences of pixels in addition to their pixel features (color and texture) while other models consider the pixel low level features only. This basic difference adds an extra degree of segmentation accuracy especially while handling noisy image data. Natural color images are particularly noisy due to the environment they were produced. Therefore, it is hard to develop a robust and faithful unsupervised technique for automatic determination of number of objects in a color image. Although there are a few existing approaches for determining the number of components in an image automatically, none of them has been found robust in all situations. As a result, we introduce *Cluster Ensembles* to get a consolidated clustering for finding the number of components in a natural color image using SNOB [2], [3] as a starting point. Using a finite/classical mixture model in color image understanding and segmentation is a popular approach though computationally intensive algorithms are a common drawback in such statistical model-based methods. Even then, finite mixture models have received significant attention from researchers in the last decade. In this work, we introduce a Maximum a priori (MAP) classifier based on a finite mixture model framework, where Expectation-Maximization algorithm has been employed for simultaneous model parameter estimation during pixel clustering. Further, the model has been extended into a multidimensional finite mixture model to enable the handling of multiple cues/features (intensity, color,

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_54,

texture, and coordinates) simultaneously. Further, Markov Random Fields (MRF) model is introduced to capture the spatial relationship among neighboring pixels in terms of its color and texture features [1]. In doing so, it can efficiently handle specific situations like recognition of different objects in an image which is usually associated with numerous artifacts such as noise and intensity inhomogeneity. So, the proposed model has some advantageous features, especially in the case of contextually dependent physical phenomena like images as it incorporates the outstanding features of the MRF model without additional computational burden.

The remainder of this paper proceeds as follows: In Section II we present an EM Algorithm Fitting MAP Classifier followed by the proposed image segmentation model using the aforesaid MAP classifier in Section III. In Section IV, we present previous work in this specific field. Section V presents experimental results demonstrating the effectiveness and accuracy of the proposed approach and a discussion on the experimental results and finally in Section VI, we present our conclusion.

II. STOCHASTIC MODEL BASED MAP CLASSIFIER

Let us assume i_x is the observation at the *x*th pixel of an image ($x = 1,....K$), and we have a Gaussian mixture model with an unknown number of constituents or components *N* where, each component has its own vector of density parameter λ_y for each feature.

So, the probabilities of the ***x***th pixel belonging to the ***y***th class label are

$$\xi_y^x = P(y \mid i_x) \tag{1}$$

where, i_x = observation at the ***x***th pixel.

The probability density function $f(i_x \mid \Gamma, \Omega)$ at any observation i_x can be expressed as:

$$f(i_x \mid \Gamma, \Omega) = \sum_{y=1}^{N} \xi_y^x \Phi(i_x \mid \lambda^y) \tag{2}$$

where, $f(\Phi(i_x \mid \lambda^y)$ is a Gaussian distribution with parameters $\lambda_y = (\mu_y, \sigma_y)$, and *N* is the number of components in the mixture.

Maximum a posteriori (MAP) introduces a prior distributionfor the parameter set that takes into account spatial information based on the Gibbs function or distribution. According to the Hammersley-Clifford theorem, Gibbs distribution takes the following form;

$$P(\Gamma) = \frac{1}{Z}\exp(-\alpha \sum V_{s_x}(\Gamma) \tag{3}$$

where, α is called the regularization parameter, the normalizing constant *Z* is called a partition function, and V_{S_x} denote clique potentials of the label configuration within the neighborhood S_x of the *x*th pixel.

An MRF is characterized by its local property whereas a GRF (Gibbs Random Field) is characterized by its global property. At this point, we need a global or joint probability distribution to represent a whole image. The Hammersly-Cliford theorem establishes the equivalence of these two types of distributions. The theorem states that any conditional distribution has a joint distribution which is Gibbs if the following conditions are satisfied;

Positivity: $\Pr(X = x) > 0$

Locality: $\Pr(X_s = x_s \mid X_t = x_t, t \neq s, t \in S)$
$= \Pr(X_s = x_s \mid X_t = x_t, t \neq s, t \in N_s)$
where, N_s is the neighborhood of site ***s***.

Homogeneity: $\Pr(X_s = x_s \mid X_t = x_t, t \neq s, t \in N_s)$ is the same for all sites ***s***.

Clique potentials of the neighborhood can be calculated as;

$$\sum V_{S_x}(\Gamma) = \sum_{l \in K_x} p(v_{x,l}) \tag{4}$$

where the $v_{x,l}$ denotes the distance between the two label vectors p^x and p^l. The function $p(v)$ must be gradually increasing and non-negative.

A posterior log-density function can be derived from Eq. (3) as;

$$P(\Gamma, \Omega \mid I) = \sum_{x=1}^{K} \log(x^i \mid \Gamma, \Omega) + \log P(\Gamma) \tag{5}$$

The *Expectation-Maximization* (EM) algorithm requires that the computation of the conditional expectation values z_y^x of the hidden variables must be at the Expectation step due to MAP estimation of the parameters $\{\xi_y^x\}$ and $\{\lambda_j\}$;

$$z_y^{x^{(t)}} = \frac{\xi_y^{x^{(t)}} \Phi(i^x \mid \Omega_y^{(t)})}{\sum_{m=1}^{N} p_m^{x^{(t)}} \Psi(i^x \mid \Gamma_y^{(t)})} \tag{6}$$

Then, maximization of the following log-likelihood corresponding to the complete data set is performed in the M-step;

Suppose,

$$U = \sum_{x=1}^{K} \sum_{y=1}^{N} z_y^{x^{(t)}} \{\log(\xi_y^x) + \log(\pi(i^x \mid \lambda^y))\}$$

Therefore,

$$MAP(\Gamma,\Omega \mid \Gamma^t\Omega^t) = U - \alpha \sum_{x=1}^{K} \sum_{l \in S_x} p(v_{x,l}) \quad (7)$$

where, t = iteration step.

For each parameter, *MAP* can be maximized independently, which yields the update equations for parameter of the component densities $\mu_y^{(t+1)}$ and $[\sigma_y^2]^{(t+1)}$.

However, maximization of the function $MAP\psi$ with respect to the label parameters ξ_y^x ψdoes not provide a closed form of update equations. In addition, the maximization procedure must also take into account the constraints $0 \le \xi_y^x \le 1$ and $\sum_{y=1}^{N} \xi_y^x = 1$. To address these difficulties, we take some new measures in the M-step of EM algorithm. We set derivative of Eq. (7) equal to zero to get the expression in the following quadratic form;

Suppose,

$$E = 4\alpha[\sum_{l \in K_x} p(v_{x,l})](\xi_y^{x^{(t+1)}})^2$$

Therefore,

$$E - 4\alpha[\sum_{l \in K_x} p(v_{x,l})\xi_y^l](\xi_y^{x^{(t+1)}}) - z_y^{x^{(t)}} = 0 \quad (8)$$

In the above equation the neighborhood K_x can include pixels with updated label parameter vectors in $(t+1)$ step, as well as pixels whose label vectors ξ^l have not yet been updated in the t step.

Now, we get two roots of the above equation as under;

Suppose, $F = \dfrac{[\sum_{l \in K_x} p(v_{x,l})]}{2[\sum_{l \in K_x} p(v_{x,l})]}$

Therefore,

$$\xi_y^{x^{(t+1)}} = F \pm \frac{\sqrt{[\sum_{l \in K_x} p(v_{x,l}) + \frac{1}{\alpha} z_y^{x^t} [\sum_{l \in K_x} p(v_{x,l})]]}}{2[\sum_{l \in K_x} p(v_{x,l})]} \quad (9)$$

From the above two roots, we select the root with positive sign + as it yields $\xi_y^x \ge 0$ which provides a straight forward update for the values of label parameters ξ_y^x of each pixel *x* at the M-step of each iteration of the EM algorithm. To satisfy the other constraints i.e., $0 \le \xi_y^x \le 1$ and $\sum_{y=1}^{N} \xi_y^x = 1$, we utilize an active set method that use Lagrange multiplies.

III PROPOSED COLOR IMAGE SEGMENTATION APPROACH

In this particular work, we have used YCrCb and Daubechies wavelet transforms as color and texture descriptors respectively. Using the combined effect of a YCrCb color model and Duabechies wavelet transform, our proposed approach is able to identify and segment different objects in an image more precisely in a meaningful manner. As a whole, our proposed segmentation approach combines a modified EM algorithm and MRF model based MAP classifier, Daubechies wavelet transform and YCrCb color space along with a robust technique for finding the number components in an image using *cluster ensembles* to implement a robust unsupervised object identification and segmentation approach that will produce accurate and faithful object identification on a wide variety of natural color images. The most significant aspect in image content understanding is feature extraction. Accuracy in object identification greatly depends on the choice of feature (intensity, color, texture, and co-ordinate) descriptor. The reason for using YCrCb color space is that the human eye is less sensitive to chrominance than luminance. Segmentation techniques can take advantage of this phenomenon and subsample the values of Cb and Cr without significant visual degradation of the basic RGB color model. The texture descriptor plays a significant role in the analysis of contextually dependent physical phenomena like image in addition to the color descriptor. This project aims to investigate and implement a more content sensitive image understanding and object identification technique that can produce reliable segmentation of different objects in color images on the basis of subtle color and texture variation and their neighborhood relationships. In order to do that we have introduced Daubechies wavelet transforms for the first time as a texture descriptor under Markovian framework for color image segmentation purpose. The Daubechies wavelet transforms has the ability to capture subtle information about the texture while other texture descriptors are vulnerable in those specific situations. Recently, approaches using MRF model and Bayesian method have provided answers to various contemporary problems in image segmentation. The application of a clustering method to image segmentation has the particular characteristics that spatial information should be taken into account. Here in addition to intensity values of pixels, the pixel location must also be used to determine the cluster to which each pixel is assigned. Conceptually, in most cases it is desired that the same cluster label be assigned to spatially adjacent pixels. In implementing these ideas, the Bayesian framework provides a natural approach. Pixel's color intensity and texture information are captured by the likelihood term in Bayesian Networks, while a prior biasing term captures the spatial location information with the help of Markov Random Fields (MRF) model.

A. Color Feature Extraction

Color is an important dimension of human visual perception that allows discrimination and recognition of visual information. Extracting and matching color features are relatively easy. In addition to that, color features have been found to be effective in color image segmentation [1], [4], [5], [6]. In this particular work, we prefer YCrCb as it has got some specific advantage in vision and graphics as mentioned earlier.

B. Texture Feature Extraction

Texture characterizes local variations of image color or intensity. There is no formal or unique definition of texture though texture based methods are commonly being used in computer vision and graphics. Each texture analysis method defines texture according to its own need. Texture can be defined as a local statistical pattern of texture primitives in observer's domain of interest. Texture is often describe as consisting of primitives that are arranged according to placement rule which gives rise to structural methods of texture analysis that explicitly attempts to recover the primitives and the replacement rules. On the other hand, statistical methods consider texture as random phenomena with identifiable local statistics. The neighborhood property is a common similarity among all the texture descriptions. So, we can define texture as an image feature which is characterized by the gray value or color pattern in a neighborhood surrounding the pixel.

We are using Wavelet transform to extract texture feature. Using Wavelet transform is a recent trend in image processing. Although, wavelet transform has made significant contribution in the several areas of image processing such as image enhancement, image compression, and image registration, its use in image segmentation is very limited. Generally, Wavelet transforms performs better than other texture descriptors like Gabor and MRSAR filters [10], [16]. Gabor filters are found efficient in capturing the strong ordered or coarser texture information while MRSAR filters works well in capturing weak ordered or finer texture details. Wavelet transform computation involves recursive filtering and sub-sampling and each level, it decomposes a 2D signal into four sub-bands, which are often referred to as LL, LH, HL, and HH (L=Low, H=High) according to their frequency characteristics. In this particular work, we use Daubechies wavelet transforms, a family of wavelet transforms discovered by Ingrid Daubechies to add more texture sensitiveness into our proposed approach. It has similar concepts like Haar wavelet but differs in how scaling functions and wavelets are defined.

C. Finding The Number of Components

Considering the pros and cons, we have selected SNOB [2], [3], an unsupervised data clustering approach which has been found more effective in handling noisy color image data that have a diverse varieties of faults and imperfections.

D. Cluster Ensembles

In general SNOB has been found more or less effective in identifying the number of components in an image but sometimes it produces spurious results due to its limitations to handle some imaging artifacts like noise, shadow, and intensity inhomogeneity. So, there is a pressing need to devise a robust unsupervised technique to find the number of components in an image automatically. An emerging approach like *Ensembles of Clusterings* can plays a significant role to augment the technique for finding a faithful and authenticated number of components in an image. Ensembles of clustering are basically a combination of multiple clusterings to get a consolidated clustering without accessing the original data and algorithms. Here, only symbolic cluster labels are mutually shared. Multiple clusterings can be produced either using a single clustering algorithm by varying the different parameter or using different clustering algorithms on same data set and parameter [11], [12]. We use SNOB as a starting point for getting multiple clusterings based on which Ensembles of Clusterings produce a consolidated clustering. The procedure we have followed for determining the number of components is as follows: produce multiple clusterings for different parameter values and select the stable solution found in the plot of the number of clusters. The convergence criteria or number of iteration of MRF model is the only parameter that we have varied to get multiple clusterings.

E: Segmentation of Image

After extraction of the six colors and texture features, we put all the features into the *Cluster Ensembles* where the SNOB acts as a starting point to determine the number of classes in the input image on the basis of these color and texture features. Now, the six original features along with the number of classes to be obtained from *Ensembles of Clusterings*, are the input of our proposed MAP classifier as described in Section 2, where clustering of pixels will be done on the basis of homogeneity in color and texture properties along with the neighborhood relationships among the neighboring pixels. As a result, regions are obtained where there is a discontinuity in either color or texture and neighborhood.

.

IV. RELATED PREVIOUS WORK

Recently, we [1] proposed a MRF model based unsupervised color image segmentation approach, which is mainly focused on textured color image segmentation using Haar wavelet transform, CIE-Luv color space, and neighborhood influences of neighboring pixels and found promising results while applied on relatively less complex images. In this work, first order neighborhood has been applied to capture the neighborhood influences among the neighboring pixels.

Kato and Pong suggested a model based color image segmentation approach that aims at combining color and texture using Markov Random Field model theory [4]. The approach relies on Bayesian classification associated with combinatorial optimization or simulated annealing where all the pixels are classified into different groups to form segments. The authors have used the perceptually uniform CIE-Luv color space and Gabor filters as color and texture descriptors respectively. It appears from their experiments that they have tested their algorithm on synthetic and simple natural images only. Experimental results reveal that this approach is able to produce coarser (most prominent objects only) segmentation on very simple images only. Further, no model parameter estimation and number of components estimation technique has been suggested by the authors which makes the approach primarily supervised in nature.

In 2003, Kato et al. proposed a similar approach using a multi-layer MRF model for unsupervised segmentation that automatically determines the number of components in the Gaussian mixture and the associated model parameters [6]. To find out the number of components in the mixture, they have introduced a mathematical term to the energy function of each layer. The paper does not provide any detail about how it works. Gaussian parameter estimation has been done by using an adaptive segmentation scheme where at every 10th iteration, they simply recomputed the mean values and covariance matrices based on the current labeling and the initial estimates at feature layers are obtained via mean shift clustering. In this paper, they have used perceptually uniform CIE-Luv color values as color features and a set of Gabor and Multi-Resolution Simultaneous Auto-Regressive (MRSAR) as texture features.

Kato and Pong suggested another stochastic model based color image segmentation method based on MRF model, which aims at combining color and texture features [5]. The theoretical frame works relies on Bayesian estimation via combinatorial optimization. This work is basically a comparative study among the segmentation results produced by color and texture features independently and in combination. The authors have tested their algorithm on a variety of color images including synthetic images and have compared their results for supervised and unsupervised segmentation using color only, texture only and combined features. No technique for automatic determination of number of components has been employed; instead they manually set the number. Using Expectation-Maximization (EM) algorithm, they have estimated the model parameters only which makes it semi-unsupervised in nature. So, their claims regarding the comparison of results between supervised and unsupervised segmentation is invalid as the algorithm they have utilized for unsupervised segmentation was not completely unsupervised in nature.

Yiming et al. [10] suggested another unsupervised color image segmentation approach based on stochastic model where they employed Minimum Message Length (MML) principle for finding the number of components in images and Maximum Likelihood Estimate (MLE) for pixel classification. This is completely a color feature based approach where neither texture feature nor has neighborhood relationship been considered.

V. DISCUSSSION ON EXPERIMENTAL RESULTS

To qualitatively evaluate the performance of the proposed approach we tested this algorithm on a wide variety of natural images comprising different number of regions or objects and get promising results but due to space constraint we would not be able to present all of them. The images are sourced from Bakerely Segmentation Dataset [11].

Figure 1 and Figure 2 show how our proposed approach takes advantage of the discriminating ability to identify and extract different objects from an image with the highest accuracy.

Regarding the proposed approach, our overall aim was to find an efficient and appropriate machine vision technique to identify and extract individual meaningful objects in textured color images. The results of the proposed algorithm have been compared with that of one of our similar previous work [1] in order to make a direct comparison of the segmentation accuracy between these approaches. Our proposed approach differs with the existing method in four major ways; 1) a second order neighborhood is employed to capture the neighborhood relationships among the neighboring pixels more efficiently, whereas in our existing method [1], first order neighborhood was employed, 3) to have better performance in texture feature extraction especially in terms of strong texture, we use Daubechies wavelet, whereas in our existing method we employed the Haar wavelet only, 3) YCrCb color space is employed to get a better human perceptual color features, whereas in our existing method, we employed CIE-Luv. Further, in our previous work, MRF model, SVFM model, Simulated Annealing, and EM algorithm were loosely coupled. Our new framework uses an integrated model to tightly couple these techniques.

The comparison of results reveals that our proposed approach outperform the existing similar approach in terms of segmentation accuracy and has higher discriminating ability to identify disjointed segments or regions.

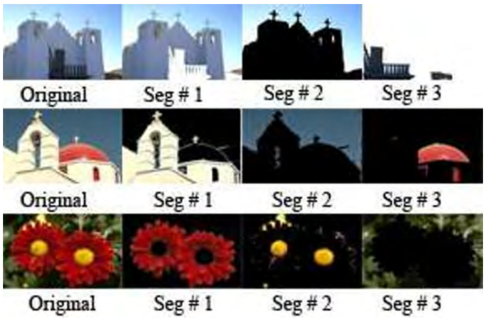


Fig. 1. Segmentation Results with our proposed approach

In Figure 2, the existing method has found 2 segments in the top image but our proposed approach able to identify 3 segments (Figure 1). If we look carefully to the image, in existing approach wooden balcony in the image is merged with the building though there is a significant color and texture difference in between them, while the proposed approach able to identify the wooden balcony as disjoint region or segments. Our approach is able to capture the subtle similarities in color and texture and able to extract or segment the region on the basis of these similarity. In Figure 1, our approach has identified 3 different objects in the middle image while the existing approach found 4 segments or objects (Figure 2). In this case, texture features plays the dominant role in the segmentation process so ignores the dissimilarity in color and considers the region as a single segment. In our proposed approach, we employ a powerful texture descriptor (Daubechies wavelet) while the existing approach employed a comparatively less effective texture descriptor (Haar wavelet). Same thing happened to the bottom image. Here our proposed approach able to identify 3 disjoints objects while on the other hand the existing approach identified 4 objects. The existing method identified 2 individual segments in the petal based on the color difference while the proposed approach able to identify the petal accurately as it is more sensitive to texture.

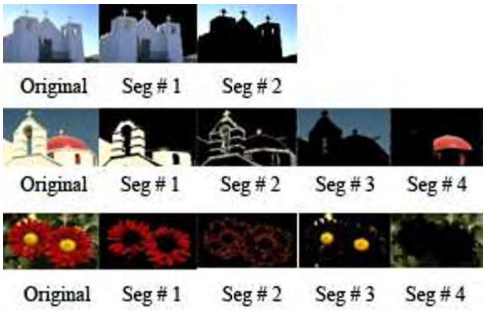


Fig. 2. Segmentation Results with the existing methods

Further, a quantitative performance evaluation has been done using a set of evaluation criteria in the form of an evaluation matrix as discussed in Section 3.5.2 in [12]. The inputs for generating the evaluation matrix is the segmentation results as obtained from our proposed approach and the existing method. The output from these evaluation criteria is a set of scores as determined by the criteria. The final evaluation matrix is generated by combining the two scores as produced by two segmentation methods. Highest score determine the best segmentation results in terms of segmentation accuracy and finding the correct number of objects in an image.

In Fig. 3, we present an evaluation matrix for quantitative comparison of results as obtained from these two approaches. The higher scores in the performance evaluation matrices further confirm that the proposed approach clearly outperforms the existing method.

Top Image:

	Proposed	Existing
MOD_LEV1	0.7944	0.6361
MOD_LEV2	0.7309	0.6589
ROS2	0.7436	0.6211

Middle Image:

	Proposed	Existing
MOD_LEV1	0.8125	0.6987
MOD_LEV2	0.7845	0.6352
ROS2	0.6574	0.6214

Bottom Image:

	Proposed	Existing
MOD_LEV1	0.7365	0.6924
MOD_LEV2	0.6428	0.5987
ROS2	0.6133	0.5587

Fig. 3. Comparative performance evaluation between the proposed and existing approach

VI. CONCLUSION

We proposed a straight-forward, robust and effective algorithm for identification and extraction of different meaningful objects in a color textured image. The evaluation of the experimental results proves the potential of the proposed approach for unsupervised object identification or extraction applications.

REFERENCES

[1] M. Islam, J. Yearwood, and P. Vamplew, "Unsupervised Color Textured Image Segmentation Using Cluster Ensembles and MRF Model", *Advances in Computer and Information Sciences and Engineering*, Springer-Verlag, 2008, ISBN: 978-1-4020- 8740-0, pp. 323-328.

[2] C.S. Wallace, D.M. Boulton, "An informal measure for classification", *Computer Journal*, Vol. 11(2), pp. 185-194.

[3] C.S. Wallace, D.L. Dow, "MML clustering of multi-state, poisson, von misses circular and gaussian distribution", *Statistics and Computing*, Vol. 10(1), Jan. 2000, pp. 73-83.

[4] Z. Kato and T.C. Pong, "A Markov Random Field Image Segmentation Model for Color Textured Images", *Image and Vision Computing*, Vol. 24(10), pp. 1103-1114.

[5] Z. Kato and T.C. Pong, "A Markov Random Field Image Segmentation Model for Combined Color and Textured Features", W. Starbek (ED), *Proceedings of ICCAIP*, Vol. 2124 of LNCS, Springer, Warsaw, pp. 547-554.

[6] G.R. Shivani, P. Manika, and D. Shukhendu, "Unsupervised Segmentation of Texture Images Using a Combination of Gabor and Wavelet Features", ICVGIP 2004, *Proceedings of the Fourth Indian Conference on Computer Vision, Graphics & Image Processing*, Kolkata, India, Vol. December 16-18, 2004, pp. 370-375.

[7] A.L.N. Fred and A.K. Jain, "Data Clustering Using Evidence Accumulation", *In the Proceedings of the 16th Intl. on Pattern recognition 'ICPR'02'*

[8] A. Strehl and J. Ghosh, "Cluster Ensemble - a knowledge reuse framework for combining multiple partitioning", *Journal of Machine Learning Research*, Vol. 2(2002), pp. 583-617.

[9] Randen and J.H. Husoy, "Filtering for Texture Classification: A Comparative Study", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. April, 1999, pp. 291-310. Analysis and Machine Intelligences, Vol. 23(8), pp. 800-810

[10] W. Yiming, Y. Xiangyu , and C.K. Luk, "Unsupervised Color Image Segmentation based on Gaussian Mixture Model", *In Proceedings of the 2003 Joint Conf. of the 4th Intl. Conf. on Information, Communications and Signal Processing*, Vol. 1(15-18 Dec. 2003), pages: 541-544.

[11] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics", *Proc. 8th Int'l Conf. Computer Vision*, July 2001, Vol. 2, pp. 416-423.

[12] Islam, M., "Unsupervised Color Image Segmentation Using Markov Random fields Model", *Master by Research Thesis*, Graduate School of Information Technology and Mathematical Sciences, University of Ballarat, Australia, 2008.

An Unsupervised Stochastic Model for Detection and Identification of Objects in Textured Color Images Using Segmentation Technique

Mofakharul Islam[1] Paul A. Watters[2]
Internet Commerce and Security Laboratory
Graduate School of Information Technology and Mathematical Sciences
University of Ballarat, Australia
mo.islam@ballarat.edu.au[1]
p.watters@ballarat.edu.au[2]

***Abstract*-The process of meaningful image object identification is the critical first step in the extraction of image information for computer vision and image understanding. The disjoint regions correspond to visually distinct objects in a scene. In this particular work, we investigate and propose a novel stochastic model based approach to implement a robust unsupervised color image content understanding technique that segments a color textured image into its constituent parts automatically and meaningfully. The aim of this work is to produce precise segmentation of different objects in a color image using color information, texture information and neighborhood relationships among neighboring image pixels in terms of their features using Markov Random Field (MRF) Model to get the maximum accuracy in segmentation. The evaluation of the results is done through comparison of the segmentation quality and accuracy with another similar existing method which demonstrates that the proposed approach outperforms the existing method by achieving better segmentation accuracy with faithful segmentation results.**

I. INTRODUCTION

Feature-based image segmentation has become an important research focus in the image segmentation community. Several authors have suggested different approaches to improve the performance and efficiency of pixel classification but none of them has been found accurate in all cases. For measuring cluster similarity and finding the numbers of clusters each approach has used its own perspective which makes it challenging to devise a robust clustering approach. For a long time, the statistics, data mining, and machine learning communities have been investigating to find such a robust approach.

Image data has some peculiar characteristics that differs entirely it from other form of data. Image data may have corrupted values due to the usual limitations or artifacts of the imaging devices. Noisy data, data sparsity, and high dimensionality of data cause difficulties in image pixel clustering. As a result, image pixel clustering becomes a harder problem than other form of data. Although there are a few existing algorithms for unsupervised color image segmentation, but all of them are based on high level features like line, edge, and area. As a result, none of them has been found to be robust in determining accurate number of components or segments.

As a rule, color image contains noisier image data. Imaging under different environments makes this problem more difficult. Therefore, it is hard to develop a robust and faithful unsupervised technique for automatic determination and segmentation of number of objects in a color image.

The presence of intensity inhomogeneities is perhaps the primary difficulty with color textured image segmentation. The reason behind these inhomogeneities caused by electro-magnetic interference of the imaging devices (camera and sensor). Apart from that, the limitations of image capturing devices lead to further difficulties in getting noise free images in some specific application areas, where security and surveillance miniature camera/sensor and artificial illumination sources are being used. A miniature camera and artificial unidirectional illumination light sources are not quite enough for homogenous illumination of an object or interest and its surroundings. As a result, captured images suffer from several varieties of imperfections including white spots due to reflectance, unclear edges, and shadows which eventually results in inhomogeneity.

Naturally, image segmentation demands well defined borders of objects as future analysis and processing is completely based on accurate identification of objects. So, there is a fundamental demand of accuracy. The segmented regions or components should not be further away from the true object than one or a few pixels. So, there is a need for an improved image segmentation technique that can segment different objects precisely.

Recently, approaches using Markov Random Fields (MRF) models and Bayesian methods have provided answers to various contemporary problems in image segmentation [1],[2], [3],[7],[5]. The theoretical framework relies on Bayesian estimation via combinatorial optimization. Different pixels represent different segmented regions in the input image. These classes are represented by multivariate Gaussian distributions.

The remainder of this paper proceeds as follows: In Section II we describe our proposed stochastic model, which includes color and texture feature descriptors, a robust tool for finding the number of components/objects and the stochastic

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_55, © Springer Science+Business Media B.V. 2010

MAP classifier that we use for image segmentation purpose. In Section III, we present previous work in this specific field. Experimental results demonstrating the accuracy and efficiency of the proposed approach are discussed in Section IV and finally in Section V, we present our conclusion and future work.

II. RELATED PREVIOUS WORK

Sanjay-Gopal and Herbert [1] proposed a novel approach for pixel labeling and gray scale image segmentation based on a Bayesian framework using a finite mixture model where an EM algorithm is used for maximum likelihood estimation of the pixel labels and the parameters of the mixture densities. The Bayesian framework provides us with a natural approach to implement the idea that the application of clustering methods to image segmentation has the particular characteristic, which takes spatial information of the pixel into consideration in addition to its intensity values. According to this formulation, a likelihood term is introduced that is based exclusively on pixel intensity information while spatial location information is captured by a prior biasing term that uses MRF model theory.

Recently, we [2] proposed a finite model based unsupervised color image segmentation approach, which is mainly focused on textured color image segmentation using Haar wavelet transform, CIE-Luv color space, and neighborhood influences of neighboring pixels and found promising results while applied on relatively less complex images.

In [3], a novel approach for segmentation of color textured images based on Markov Random Fields has been proposed where they use CIE-Luv color values as color features and a set of Gabor filters as texture features. The proposed approach is supervised in nature. As such, no model parameter estimation technique has been used and the number of components was selected manually. In 2003, Kato et al. [7] proposed an algorithm using a multi-layer MRF model for unsupervised segmentation that automatically estimates the number of components in the Gaussian mixture and the associated model parameters. To find out the number of components in the mixture, they introduced a mathematical term to the energy function of each layer. Gaussian parameter estimation has been done by using an adaptive segmentation scheme where at every 10th iteration, they simply recomputed the mean values and covariance matrices based on the current labeling and the initial estimates at feature layers are obtained via mean shift clustering. In this paper, they have used perceptually uniform CIE-Luv color values as color feature and set of Gabor and MRSAR filters as texture feature. Further, Kato and Pong [5] proposed another similar algorithm where they used CIE-Luv color values as color feature and a set of Gabor filters as texture feature. They utilized the EM algorithm for estimation of the model parameter. They set the number of components (pixel class) manually. In another approach [8], an unsupervised segmentation algorithm which uses Gaussian MRF model for color textures. Their model characterizes a texture in terms of spatial interaction within each color plane and interaction between different color planes. They proposed an Agglomerative Hierarchical Clustering for pixel segmentation. The model parameter has been estimated using a Maximum pseudolikelihood scheme but no automatic technique has been used for estimating the number of mixture components.

III. PROPOSED IMAGE SEGMENTATION MODEL

Our proposed image segmentation model composed of following three major sequential phases;

1. Image feature extraction
2. Determination of number of components in the image using a robust number of object/component finding tool
3. Pixel classification using a stochastic model based MAP classifier

A Image Feature Extraction

Color textured image has different features like intensity, color, texture and coordinates. In our project, we have used color and texture features for the segmentation purpose. To segment an image into objects, nine features are extracted from the image. Three features are color features and the rest are texture features. CIE-Lab color space has been chosen for color feature extraction considering its several advantages as discussed in [11]. To obtain the other six features, we apply the Haar wavelet transform to the *L* components (number of pixels) of the image to get three features and Gabor filters under three different orientations for the rest three features.

B Color Feature Extraction

Choice of a suitable color space is one of the main aspects of color feature extraction. Most of the color spaces are three dimensional where different dimensions represent the different color components. The most common color space in computer graphics is RGB. Unfortunately, this 3D color space does not correspond to equal perception of color dissimilarity. As a result, alternative color spaces are generated by transformation of the RGB color space. The color spaces like HSV, CIE-Luv and CIE-Lab can be generated by non-linear transformation of the RGB space. We prefer CIE-Lab as it represents the three characteristics (hue, lightness and saturation) that characterize the perceptually uniform color efficiently.

C Texture Feature Extraction

The texture features are constructed in such a way that similar texture represents similar intensities. As a result, a well determined value with some variance is assigned to the pixels with a given texture. The image textures are being extracted through the wavelet transform as already mentioned earlier. Wavelet transform computation involves recursive filtering and subsampling; and each level, it decomposes a 2D signal into four sub-bands, which are often referred to as LL, LH, HL, and HH (L=Low, H=High) according to their frequency characteristics [9].

Degraded images like images from surveillance camera have much noise due to the inherent technical difficulty of the device. As such, we need some robust texture descriptors to extract our texture features. In addition to the Haar wavelet transform, we employ multi-channel filtering approach where the channels are represented by a bank of real-valued, even-symmetric Gabor filters to extract texture features as it has been found to be effective in capturing stronger texture components.

We extract three texture features using Haar wavelet transform. After a one-level wavelet transforms, a 4*x*4 block is decomposed into 4 frequency bands, each band containing a 2*x*2 matrix of coefficients. Suppose the coefficients in the HL band are $C_{k+i}, C_{k,l+1}, C_{k+1,l}, C_{(k+1,l+1)}$. Then the feature of the block in the HL band is computed as:

$$f = \left(\frac{1}{4}\sum_{i=0}^{1}\sum_{j=0}^{1} C^2_{k+i,l+j}\right)^{\frac{1}{2}} \tag{1}$$

The reason for choosing Haar wavelet is that it has better reflection of texture properties where the coefficients in different frequency bands signal variations in different directions, such as horizontal, vertical, and diagonal [4]. In addition to that, Haar transform require less computation compared to other wavelet transform with longer filters. We extract three texture features using Haar wavelet transform.

Rest of the three texture features are being extracted using Gabor filters under different orientations. Different orientations can be obtained by rotating the coordinate system. In our tests, we used three orientations: 0^o ; 60^o ; 135^o.

D Determination of number of classes

We have tried several data classification approaches to identify the number of components automatically but most of them has been found inappropriate for handling image data properly and need major adjustment through a huge series of trial and error. In that perspective, AutoClass C [10], an unsupervised Bayesian classification approach from NASA has been found effective in automatic determination of the number of classes in the input images on the basis of supplied color and texture features.

E Stochastic Model Based MAP Classifier

Let $X_i = (x_1^i, x_2^i, \ldots x_n^i)$ denote the observation of *i*th pixel of the image, where x^i ***n*** represents pixel values in the *n*th dimensional plane. So, the probabilities of the *i*th pixel belonging to the ***j***th class in the *n*th plane are given as:

$$r_j^i = P(j \mid x_n^i) \tag{2}$$

The objective function of our proposed MAP classifier is defined as;

$$MAP(\Phi, \Omega \mid \Phi^t \Omega^t) = U - \alpha \sum_{i=1}^{N} \sum_{h \in N_i} q(l_{i,h}) \tag{3}$$

Subject to $0 \le r_j^i \le 1$ and $\sum_{j=1}^{K} r_j^i = 1$,

Where, $U = \sum_{i=1}^{N} \sum_{j=1}^{K} r_j^{i^{(t)}} \{\log(r_j^i) + \log(\sigma(x_n^i \mid \theta^j))\}$,

t=iteration step, Φ =set of probability vector $\{r^1, r^2, \ldots r^n\}$ for pixel ***i***, Ω = set of component parameters $\{\theta^1, \theta^2, \ldots \theta^K\}$, α =regularization parameter, $q_{(l_{i,h})}$ = distance between two adjacent label vectors r^i and r^h, r_j^i = label parameter, x_n^i =***i***th value of pixel in the ***n***th plane, θ^j = parameters $\{\mu_j, \sigma_j\}$, ***K*** = number of components, and ***N*** = number of pixels. The algorithm iteratively maximizes the objective function (3) using the following equation (4 – 8). The conditional expectation value of the hidden variables $\rho_j^{i^{(t)}}$ is derived as follows;

$$\rho_j^{i^{(t)}} = \frac{r_j^{i^{(t)}} \delta(x_n^i \mid \theta^{j^{(t)}})}{\sum_{k=1}^{K} r_k^{i^{(t)}} \delta(x_n^i \mid \theta^{k^{(t)}})} \tag{4}$$

The density function $f(x_n^i \mid \Phi, \Omega)$ at an observation x_n^i is defined as;

$$f(x_n^i \mid \Phi, \Omega) = \sum_{j=1}^{K} r_j^i \delta(x_n^i \mid \theta^j) \tag{5}$$

Where, $\delta(x_n^i \mid \theta^j)$ is a Gaussian distribution with parameters $\theta^j = \{\mu_j, \sigma_j\}$.

In order to incorporate the neighborhood influences among the neighborhood pixels, *Maximum a priori* (MAP) introduces a prior distribution for the parameter set - that takes into account spatial information based on the Markov Random Fields (MRF) Model theory. An MRF is characterized by its local property. At this point, we need a global or joint probability distribution to represent a whole image. The Hammersly-Cliford theorem establishes the equivalence of these two types of distributions. The theorem states that any conditional distribution has a joint distribution which is Gibbs if the following conditions are satisfied;

Positivity: $\Pr(X = x) > 0$

Locality: $\Pr(X_s = x_s \mid X_t = x_t, t \ne s, t \in S)$
$= \Pr(X_s = x_s \mid X_t = x_t, t \in N_s)$, where
N_s is the neighborhood of sites **s.**

Homogeneity: $\Pr(X_s = x_s \mid X_t = x_t, t \in N_s)$ is the same for all sites ***s***.

The following Gibbs function takes into account spatial information and express it in the form of prior distribution $p(\Phi)$;

$$p(\Phi) = \frac{1}{Z}\exp(-U(\Phi)) \qquad (6)$$

where, $U(\Phi) = \alpha\sum_{i=1}^{N} C_{N_i}(\Phi)$, Z is a normalizing constant, ® is a regularization parameter, $C_{N_i}(\Phi)$ denotes the clique potential function of the pixel label vectors $\{p^h\}$ within the neighborhood N_i of the ***i***th pixel which can be defined as ;

$$C_{N_i}(\Phi) = \sum_{h \in N_i} q(l_{i,h}) \qquad (7)$$

Where $l_{i,h} = |r^i - r^h|^2 = \sum_{j=1}^{K}(r_j^i - r_j^h)^2$ and N_i is the set containing pixels that are either horizontally or vertically adjacent to the pixel ***i***. Here, we have adopted the function $q(l) = (1 + l^{-1})^{-1}$ from [17], which is more robust to outliers.

Now, based on the above prior density, a posteriori log-density function can be defined as;

$$p(\Phi, \Omega) = \sum_{i=1}^{N} \log f(x_n^i \mid \Phi, \Omega) + \log p(\Phi) \qquad (8)$$

To estimate the parameter r_j^i and θ^h for *MAP*, the Expectation-Maximization algorithm needs the conditional expectation values ρ_j^i as defined in Eqn. (4) at the Expectation step while Maximization step maximizes the objective function of our MAP classifier as defined in Eqn. (3).

However, maximization of the function *MAP* with respect to the label parameters r_j^i does not provide a closed form of update equations. In addition, the maximization procedure must also take into account the constraints $0 \le r_j^i \le 1$ and $\sum_{j=1}^{K} r_j^i = 1$.

To address these difficulties, we take some new measures in the M-step of EM algorithm. We set derivative of Eq. (8) equal to zero to get the expression in the following quadratic form;

Suppose,

$$E = 4\alpha[\sum_{h \in N_i} q(l_{i,h})](r_j^{i^{(t+1)}})^2$$

Therefore,

$$E = 4\alpha[\sum_{h \in N_i} q(l_{i,h}) r_j^i](r_j^{i^{(t+1)}}) - \rho_j^{i^{(t)}} = 0 \qquad (9)$$

In the above equation the neighborhood N_i can include pixels with updated label parameter vectors in $(t+1)$ step, as well as pixels whose label vectors r^h have not yet been updated in the ***t*** step.

Now, we get two roots of the above equation as under;

Suppose, $F = \dfrac{[\sum_{h \in N_i} q(l_{i,h})]}{2[\sum_{h \in N_i} q(l_{i,h})]}$

Therefore,

$$r_j^{i^{(t+1)}} = F \pm \frac{\sqrt{[\sum_{h \in N_i} q(l_{i,h}) + \frac{1}{\alpha}\rho_j^i[\sum_{h \in N_i} q(l_{i,h})]]}}{2[\sum_{h \in N_i} q(l_{i,h})]} \qquad (10)$$

From the above two roots, we select the root with positive sign + as it yields $r_j^i \ge 0$ which provides a straight forward update for the values of label parameters r_j^i of each pixel *i* at the M-step of each iteration of the EM algorithm. To satisfy the other constraints i.e., $0 \le r_j^i \le 1$ and $\sum_{j=1}^{K} r_j^i = 1$, different well studied mathematical approaches are available. In this particular work, we utilize an active set method that use Lagrange multiplies.

IV. DISCUSSSION ON EXPERIMENTAL RESULTS

Recognition of a particular object in an image is a task that is carried out as pixel classification, which in many cases follows recognition of the individual object on the image. The most significant aspect in pixel classification is feature extraction. Feature extraction is a process of measurement that performs the main role in the process of object recognition. The performance at segmentation depends strongly on the choice of feature extraction.

In this work, our overall aim was to find an efficient and appropriate machine vision technique to identify individual objects in color textured images, which is capable to take into account the subtle variations in color and texture and produce precise segmentation.

A series of image segmentation experiments on a wide variety of color textured images have been carried out to evaluate and compare the effectiveness of the proposed approach. Since the main contribution of our work is development of a new stochastic model based image segmentation approach, we compared our proposed approach with one of our earlier work that employs a stochastic approach based spatially variant finite mixture model [2]. Our proposed approach differs with the existing method in four major ways; 1) we have employed a new robust technique [10] to find out the number of components from image automatically, which has been found to be robust in degraded situation, 2) a second order neighborhood is employed to capture the neighborhood relationships among the neighboring pixels more efficiently, whereas in our existing method [2], first order neighborhood was employed, 3) to have better performance in texture feature extraction especially in terms of strong ordered texture, we use Gabor filters in addition to the Haar wavelet filter, whereas in our existing method we employed the Haar wavelet only, 4) CIE-Lab color space is employed to get a better human perceptual color features, whereas in our existing method, we employed CIE-Lab. Further, in our previous work, MRF model, SVFM model, Simulated Annealing, and EM algorithm were loosely coupled. Our new framework uses an integrated model to tightly couple these techniques.

We have sourced our images from Berkeley image dataset [12]. In Figure 1, we observe that, our proposed approach is able to detect 3 segments in all the experimental images and segments the objects accordingly in a precise manner. It is able to handle usual imaging artifacts like intensity inhomogeneity and noises. Here, neighborhood relationships and texture play a crucial role in detecting and segmenting the objects. In the top image, color of the bottom part of the bear's body is quite different than that of the top part but continuity in the neighborhood and textures eventually influence these regions to become members of a single region. On the other hand, in Figure 2, we observe that the existing approach found 4 segments instead of 3. Top and bottom part of the bear's body are identified as separate objects due to poor neighborhood relationships among the neighboring pixels and less sensitivity of the texture features. Further, the sandy beach and the sea have almost similar color but identified as separate objects due to the neighborhood influence and texture sensitivity of the proposed approach. We have seen similar results in the middle and bottom images with the proposed approach, where, horses are identified accurately as separate objects. Grassy land and the yellowish vegetation on top of that are identified as separate objects while on the other hand, we observe that the existing approach again found 4 segments in the middle image. Here, we observe over segmentation in the horse objects due to discontinuity in color ignoring the similarity in texture. In the bottom image, we observe the similar results with the existing approach though it yields 3 segments like the proposed approach. Here, white regions in the forehead and front right leg merged with the background as they find better similarity in color. Similarity in texture and the neighborhood relationships among the neighboring pixels couldn't influence them to remain in the same region.

Experimental results show that our proposed approach can achieve better segmentation results utilizing the spatial information of pixels more efficiently than our earlier approach [2].

However, the main contribution of this paper focuses on how spatial information can be efficiently used by a stochastic model, and concerns only the fundamental mechanism of the model. For simplicity, degraded images like images from surveillance camera, aerial images, radar images, and images in artificial environments are not discussed in this paper. To handle those images, more complicated segmentation schemes are needed, which may further improve the clustering performance.

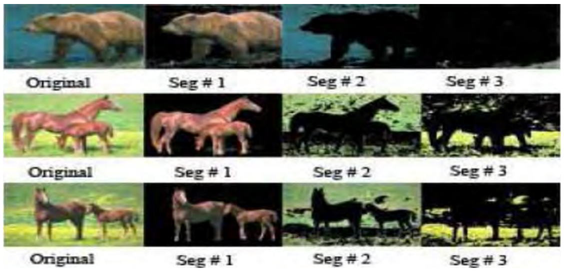


Fig. 1. Segmentation Results with the proposed approach

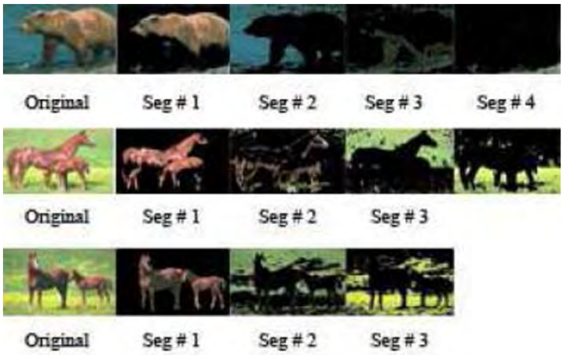


Fig. 2. Segmentation Results with the existing approach

Further, a quantitative performance evaluation has been done using a set of evaluation criteria in the form of an evaluation matrix as discussed in Section 3.5.2 in [16]. The inputs for generating the evaluation matrix is the segmentation results as obtained from our proposed approach and the existing method. The output from these evaluation criteria is a set of scores as determined by the criteria. The final evaluation matrix is generated by combining the two scores as produced by two segmentation methods. Highest score determine the best segmentation results in terms of segmentation accuracy and finding the correct number of objects in an image.

In Fig. 3, we present an evaluation matrix for quantitative comparison of results as obtained from these two approaches. The higher scores in the performance evaluation matrices further confirm that the proposed approach clearly outperforms the existing method.

Top Image:

	Proposed	**Existing**
MOD_LEV1	0.8745	0.6878
MOD_LEV2	0.7925	0.6394
ROS2	0.8436	0.5296

Middle Image:

	Proposed	Existing
MOD_LEV1	0.8421	0.7858
MOD_LEV2	0.8134	0.7294
ROS2	0.6913	0.6301

Bottom Image:

	Proposed	Existing
MOD_LEV1	0.7952	0.6541
MOD_LEV2	0.7489	0.6132
ROS2	0.6824	0.5964

Fig. 3. Comparative performance evaluation between the proposed and existing approach

V. CONCLUSION

Our aim was to propose and implement a new stochastic model based unsupervised segmentation approach for color image that can successfully segment color images with each and every details present in color and texture. We have applied our approach on a wide variety of natural color images and compared our results with one of our existing similar method and have got better results in terms of reliability and accuracy. Future work will focus on improvement of this approach to make it enable in handling degraded images like surveillance camera, aerial images, radar images, and images in artificial environments.

Further, this approach could be applied to a wide variety of color images from other application domains as well. Comparison of experimental results in terms of efficiency, robustness and segmentation quality could be done with other stochastic model based approaches. The ability of discriminating minor variations in color and texture also makes the proposed approach more appropriate in some other specific application domains like material sciences, mining, metallurgy, bio-medical, bio-informatics etc. Further, efficiency of Haar wavelet transform could be compared with other texture descriptors like Gabor and MRSAR.

Automatically segmenting objects in natural images is an important first step in biometrics systems that could be more widely used in surveillance operations at sensitive locations, such as airports. In this application area, multiple objects must be extracted in real-time, identified as faces, and then matched in a 1-1 matching against features stored in templates during enrolment [13]. Recent work had indicated that even if segmentation is possible in real-time at sufficient resolution to enable recognition, scaling-up face matching to the size of an entire nation would be difficult [14]. However, the development of new statistical frameworks such as the concept of misidentified risk [15] - will assist in the development of scalable 1- many and many-many pattern recognition techniques which would logically sit on top of high-performance (and highly-accurate) facial segmentation routines, as identified in this paper.

REFERENCES

[1] Sanjay-Gopal, S., and Herbert, T. J., "Bayesian Pixel Classification Using Spatially Variant Finite Mixtures and the Generalized EM Algorithm" *IEEE Trans. On Image Processing*, Vol. 7, Issue No. 7, July 19

[2] Islam M., Yearwood J., and Vamplew P., "Unsupervised Color Textured Image Segmentation Using Cluster Ensembles and MRF Model", *Advances in Computer and Information Sciences and Engineering*, Springer-Verlag, 2008, ISBN: 978-1-4020-8740-0, pp. 323-328.

[3] Kato Z., Pong T. C., "A Markov Random Field Image Segmentation Model for Combined Color and Textured Features" W. Skarbek (ED), *Proceedings of ICCAIP*, Vol. 2124 of LNCS, Springer, Warsaw, 2001, pp. 547-554

[4] Unser, M., "Classification and Segmentation Using Wavelet Frames" *IEEE Trans. Image Processing*, Vol. 4, No. 11, pp. 1549-1560, 1995.

[5] Kato Z., Pong T. C., *"A Markov Random Field Image Segmentation Model for Color Textured Images"* Image and Vision Computing, Vol. 24, No.10, pp. 1103– 1114, October 2006.

[6] Green, P. J., "Bayesian Reconstruction from Emission Tomography Data Using a Modified EM Algorithm" IEEE Trans. On Medical Imaging, Vol. 9, No. 1, pp. 84-93, 1990

[7] Z. Kato, T.C. Pong, and G. Song, "Unsupervised segmentation of color textured images using a multi-layer MRF model", *Proceedings of ICIP*, Vol. (1), Spain, (2003) pp. 961-964.

[8] D.K. Panjwani and G. Healy, "Markov Random Field Models for Unsupervised Segmentation of Textured Color Images", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 17(10).

[9] Chang, T., and Jay Kuo, C. C., "Texture analysis and classification with tree-structured wavelet transform" *IEEE Trans. Image Proc*., Vol. 2, No. 4, pp. 429-441, 1993

[10] Cheeseman, P., Stutz, J., *"Bayesian Classification(AutoClass): Theory and Results"* RIACS, NASA.

[11] Ihaka R., *"Colour for Presentation Graphics"*, Proceedings on the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)

[12] Martin D., Fowlkes C., Tal D., and Malik J., " Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics", *Proc. 8th Int'l* Conf. Computer Vision, July 2001, Vol. 2, pp. 416-423.

[13] Ho, W.H., Watters, P.A., and Verity, D., "Are younger people more difficult to identify or just a peer-to-peer effect?", *Proceedings of the 12th International Conference on Computer Analysis of Images and Patterns*, Vienna, Austria, 2007

[14] Ho, W.H., Watters, P.A., and Verity, D., "Robustness of the new owner-tester approach for face recognition experiments.", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* – Biometrics Workshop, Minneapolis, USA., 2007

[15] Ho, W.H., and Watters, P.A. "A New Performance Evaluation Method for Face Identification Regression on Misidentified Risk.", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA., 2007.

[16] Islam, M., "Unsupervised Color Image Segmentation Using Markov Random fields Model", *Master by Research Thesis*, Graduate School of Information Technology and Mathematical Sciences, University of Ballarat, Australia, 2008.

Automated Classification of Sunspot Groups with Support Vector Machines

Mehmmood A. Abd

Electrical And Computer Engineering Department, University Of Windsor, ON, Canada

Sarab F. Majed

Mathematic Department, University Of Western Ontario, ON, Canada

V. Zharkova

Department of Computing and Mathematics, Bradford University, UK.

***Abstract*—A new effective technique is presented for automatic classification of sunspot groups on full disk white light (WL) solar images. This technique is implemented on images taken from the Solar Oscillations Investigation Michelson Doppler image (SOI/MDI) aboard the Solar Heliospheric observatory (SOHO). The technique focuses on employing Support Vector Machines (SVMs) as effective classification tool. In addition to applying SVMs the problem of extracting sunspots and sunspot groups from solar image is solved in an efficient and different way from the ones previously employed. This technique proceeds in several consequence phases. The first phase involves solar disk image extracting. The second phase involves Binarization and smoothing of an extracted solar image disk. The third phase involves unsupervised segmentation of sunspots groups. This phase consists of two subphases: a) extracting the spots and, b) combining together the sunspots that belong to the same group. The fourth phase involves attributes extraction of each sunspots group. The final phase is the classification phase using SVMs, by applying a one–against–all technique to classify the sunspots groups. The proposed technique has been tested using different sets of sunspots achieving approximately 88.56% recognition rate for Zurich classification of sunspot groups.**

Index Term—Solar Disk, Sunspots Groups, Support Vector Machines, Unsupervised Segmentation.

I. INTRODUCTION

Sunspots have been a subject of interest to scientists and researchers for many years. A sunspot is a region in the sun's photosphere that is darker by comparison with its photosphere background and has intense magnetic field. A sunspot may consist of one or more umbra which often are surrounded by a less dark area called penumbra. Sunspot numbers increase and decrease with an irregular cycle with a length of approximately 11 years [1]. Sunspots can be observed in the visible continuous spectrum, also known as " white light '. Larger sunspots can be observed by using Call K1 absorption line image and Hα and Call K3 absorption line images . Sunspot shapes are varied from individual spots to a group of spots. Sunspot groups can have infinite different formations and size, ranging from solo spots to the giant groups of sunspots with complex structures [1]. Despite such complexity and diversity, astronomers have been able to classify the sunspots classes based on The McIntosh Sunspot Classification Scheme and modified Zurich Scheme [2]. A sunspot is recognized by its two main features ; the central dark features 'umbra'; and the outer lighter area which is called 'penumbra' (See Figure 1).Therefore,most computerized recognition techniques rely on these properties .

Many manual sunsopts in different formats are produced all over the world , such as the Meudon Observatory in France ,the Locarno Solar Observatory in Switzerland ,the Mmount Wilson Observatory in the United States Of America and many others . Zurich classification is based on the Locarno Catalogue which is considered a primary indicator of solar activity [2][3]. The Zurich Classification is categories into seven classes of sunspots groups numbered from A to H [2] .(see Figure 2 and Table 1). Because of the effect of the solar activity on the Earth's atmosphere, human activity, businesses and even climate, the automated detection and verification techniques become essential for a reliable forecast of the future effects on human lives and businesses. This requires stringent requirments on the classification accuracy of any automated recognition system. Many researchers have been developing automated reliable recognition systems with varying recognition rate and techniques [3].

Fig.. 1. A high-resolution SOHO/MDI image showing sunspots with umbra (black spots) and penumbra (dark gray) above the background (gray).

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*, DOI 10.1007/978-90-481-9151-2_56, © Springer Science+Business Media B.V. 2010

TABLE 1
SUNSPOT CLASS DESCRIPTION

Class	Description
A	Unipolar, no penumbra, length < 3
B	Bi-polar, no penumbra, length ≥3
C	Bi-polar, with penumbra on spots of one polarity only, usually the spot at one end of elongated group.
D	bi-polar, penumbra on spots of both polarities, length ≤ 100
E	bi-polar, penumbra on spots of both polarities, length > 100 and ≤ 150
F	bi-polar, penumbra on spots of both polarities, length > 15
H	Unipolar, with penumbra, principal spot is usually the leading spot remaining from an old bi-polar group

In this paper a new recognition technique is presented based using one–against-all SVMs.

The classification system receives its input from reliable segmentation phase that employed the unsupervised segmentation method [4]. The objectives of this system is to a) improve the classification rate of sunspot groups classification,b) improve segmentation techniques of sunspots groups, c) employ more reliable attributes to enhance the classification reliability.
The proposed methodology is described by the following processing phases :

1. Preprocessing
2. Segmentation technique
3. Feature extraction
4. Multi–Class SVMs Classification

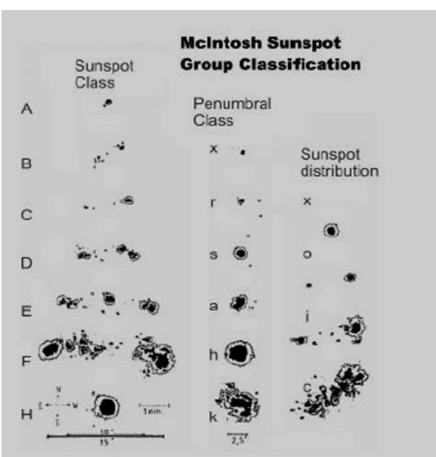


Fig.2. The Modified Zurich Sunspot Classification System developed by Patrick McIntosh [2].

The remainder of this paper is organized as follows. Section II describes preprocessing, Section III presents the segmentation technique, Section IV presents feature extraction technique, Section V presents multi-class SVMs Classification technique, Section VI discusses the results obtained from the previous sections and summarizes the conclusions .

II. PREPROCESSING

Preprocessing consists of image acquisitions, edge detection, noise removing and feature enhancing, and Binarization. The input images are taken from the solar Oscillations Investigation Michelson Doppler Imager (SOI/MDI) aboard of the solar Heliospheric Observatory (SOHO)[11]. Figure 3 shows an example of the input white light full disc image .The image has dimensions of 1024x1024 pixels with 16 different intensities. The resolution is 2.6 arces/pixel. Edge detection, noise removing and Binarization are curried out as that reported by Sarab *et al* [4].

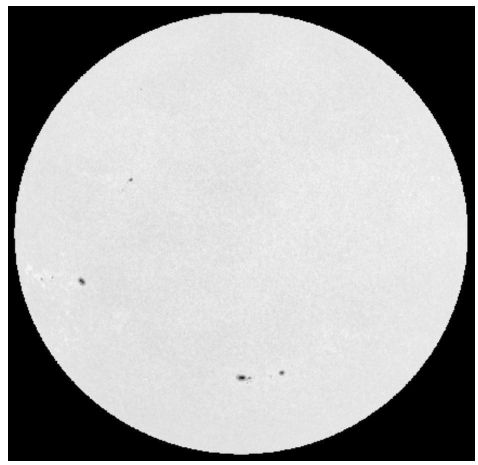


Fig. 3. Full disk image obtained from SOHO/MID on 1/1/2000.

Image binarization converts an image into black and white. In this study a threshold value, *t,* is determined automatically using Otsu's method [12] .The produced image is a binary image with white sunspots and black background. Figure 4 illustrates the output that will be feed to the next phase /segmentation phase.

Morphological processing facilitates the sunspot groups' attributes extracting. Meaningful shape information of geometrical objects, measuring typical characteristics area, bounding box, centroid coordinate, and the length of a perimeter, is computed. Finding sunspot coordinates (x,y) and area in pixels ,eases a determination of the sunspots locations in relation to the solar center . Moreover, area of any sunspot is utilized to estimate the area in heliographic coordinates. These attributes are employed later to classify each group belong to which class. Next section presents sunspots/ sunspot groups detection and grouping then extracting.

III. SEGMENTATION TECHNIQUE

In this paper an unsupervised segmentation method was employed based on the similarity measure. The sunspot's is extracted /cut from a full disk solar disk image by computing the bounding rectangle (box) that contains it.

Fig. 4. Black and white image produced from the preprocessing phase.

Then an iterative algorithm is employed to estimate the maximum bounding box that will be re-adjusted iteratively, in order to combine the group of sunspots together. The images used are WL of size N_r x N_c with intensities given S(i,j) ={Sij: 1 <= i <= Nr , 1<=j <=N_c } the produced image from the previous step is masked so that the background is black and sunspots are white. The bounding box is represented by the vector *V* where V=(x,y,xw,yw) ,and (x,y) coordinates of the left upper corner and (xw,yw) are the x-width and y-width, respectively. Then the other corners are represented by (x,y+yw) (for the upper right corner), (x+xw ,y) (for the lower left corner)and (x ,y+yw) (the lower right corner).

The algorithm starts by checking the size of each bounding box to find the one that contains adjacent sunspots. This technique starts by choosing a seed as the maximum sunspot area with the maximum bounding box. Then the maximum bounding box will be adjusted to contain the neighbor boxes. The segmentation algorithm can be written as follows. If the initial bounding box has vector $V_0=(x_0,y_0,xw_0,yw_0)$, and the neighbor box has vector , $Vr=(x_r,y_r,xw_r,yw_r)$, then the adjustable box has the vector , Vr=(x,y,xw,yw), where x=*MIN*(x_0,x_r) , and y=*MIN*(y_0,y_r), then xw=*MAX* $(x_r+xw-x,y_r+xw_r,-x)$,
yw=*MAX* (y_r+yw_r-y,y_r+yw_r-y).

The above algorithm is repeated until it finds all spots contained in each group and fits them into one box. Then the area that contains the sunspot group is extracted from the main image and saved as a separate image that represents one sunspot group. Figure 5 shows the segmentation results with four different sunspot group images where A represents the pores and groups B, C, and D respectively [2].

IV. FEATURE EXTRACTION

In this phase the attributes of each sunspots groups are extracted and formed into a vector of N (7 x 1) based on the Modified Zurich classification attributes. After extracting all the attributes the vector is fed to the SVMs classifier [6], [5]. The features consist of a number of sunspots/umbra, size, penumbra, penumbra's position, bipolar, unipolar, and major longitudinal axis. The seven classes of modified Zurich classification are used as standard classes in this paper. For further details see [2]. The classes are numbered as A, B, C, D, E, F, and H. To make the classification more reliable the class letters are converted to numbers from 1 to 7. Thus, the class A is represented by one (1) and B by (2) and so on. The attributes of each sunspot or sunspots groups are formed into vector and fed to the SVMs. Multi-class SVMs are employed in this paper.

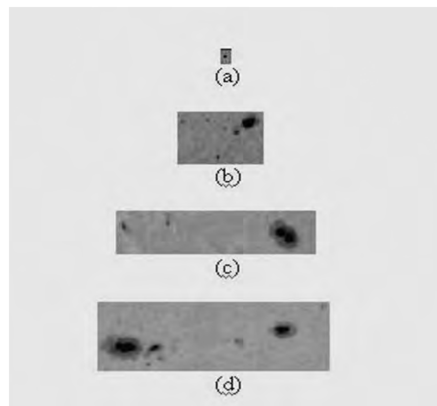


Fig. 5. Sample of sunspot groups segmentation technique applied on WL image disk SOHO/MID on 25/1/2005.

V. MULTI-CLASS SVMS CLASSIFICATION

SVMs are basically binary classifiers and, thus, it is not straightforward to turn them into multi-class (N-class) recognition systems. There are several methods invented to construct a multi–class SVM. The most typical method is to construct N SVMs, each of which classifies one class against all the other classes. This method is commonly called one-against-all (1-v-all) [7],[10]. The second method that can be used is to combine all possible two-class classifiers, where for an N-class problem *N(N-1)/2* classifiers must be constructed. This is commonly referred to as one-against-one (1-v-1). Decision Directed Acyclic Graph (DDAG) and Max Win Algorithm (MWA) are also used to construct multi-class SVMs based on (1-v-1). Further details can be found in [6],[8]. The one-against-all method is used in this study. N SVMs are constructed, where N is the number of classes. For i=1,2…N, the ith SVM is trained with all the samples in the i^{th} class considered as positive examples and the samples from all other classes considered as negative examples. Given $\{(x_1, y_1), (x_2, y_2),\ldots,(x_k, y_k)\}$ as the training data set, where $x_j \in R^n$ are the n-dimensional samples (feature vectors) and $y_j \in \{1,2,\ldots N\}$ are the corresponding class labels j=1,2,…,k, the i^{th} SVM solves the following problem [6], [9]:

$$\begin{aligned}&\operatorname{Min}_{w^i,b^i,\xi^i} \left\{\frac{1}{2}\left(w^i\right)^T w^i + C\sum_{j=1}^{k} \xi_j^i\right\}\\ &\left(w^i\right)^T \Phi\left(x_j\right) + b^i \geq 1-\xi_j^i, \text{ if } y_j = i\\ &\left(w^i\right)^T \Phi\left(x_j\right) + b^i \geq 1+\xi_j^i, \text{ if } y_j \neq i\\ &\xi_j^i \geq 0,\ j = 1,\ldots\ldots .k\end{aligned} \tag{1}$$

where 1/2 $(W^i)^T.W^i$ is the regularization term (objective function),

$$C \sum_{j=1}^{k} \xi_j^i \tag{2}$$

is a cost term (training error) used to constrain the solution in cases of non-separable data (ξ are 'slack' variables introduced for this purpose [5]), function Φ is used to map the training data into a higher dimensional space, and C is a penalty factor set by the user.

The above equation will be solved for each class $i = 1,2,..,N$, where the attempt is made to balance between the regularization term and the training errors. Thus, there are N decision functions

$$\begin{aligned}&\left(w^1\right)^T \Phi(x) + b^1\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ &\left(w^N\right)^T \Phi(x) + b^N\end{aligned} \tag{3}$$

and x is said to belong to the class which gives the largest value for the decision function, i.e.,

$$\text{class of } x \equiv \operatorname{argmax}_{i=1,\ldots,N}\left(\left(w^i\right)^T \Phi(x) + b^i\right) \tag{4}$$

In practice the dual problem, formulated in terms of Lagrange multipliers, is solved, i.e., maximize the follows

$$\begin{aligned}&\sum_i a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j \Phi(x_i)\cdot\Phi(x_j)\\ &0 \leq a_i \leq C\\ &\sum_i a_i y_i\end{aligned} \tag{5}$$

equation (5) has the following solution

$$w = \sum_{i=1}^{N_S} a_i y_i \Phi(x_i) \tag{6}$$

Where N_s is the number of support vectors and a_i Lagragian multiplier for any training point [5].

Replacing the inner product between Φ in (5) with a function simplifies things, where in kernel functions are introduced. In this work, Gaussian Radial Basis Functions (GRBF) are employed as kernel functions:

$$K(x_i, x_j) = e^{\left(-\gamma\left\| x_i - y_j\right\|^2\right)}, \gamma > 0 \tag{7}$$

The linear kernel is a special case of GRBF as Kreethi and Lin have shown in [9]. Sigmoid kernels produce results generally comparable to GRBF kernels. However, GRBF has less numerical complexities in comparison to other kernels [7]. GRBF has a control parameter, γ, which along with the cost factor C are the two tunable parameters in the SVM optimization problem. As suggested in [9], a useful way to determine these is to perform grid-search with cross-validation. A two-stage procedure can also be followed, where after identifying a good region in the (C, γ) space, a finer search can be conducted which may improve the results somewhat in some cases. The best parameters thus determined are used in the classification.

VI. RESULTS

The set of images are collected form SOHO /MID and EGSO catalogue. In this study these images are used as database, for training and testing, in this study. Twenty images were used for each class of sunspots. The images are divided into two groups of ten images.

The attributes are extracted from the images and formatted into two text files. One of the text files is used as a training resource and the other as a test resource. In the final testing stage, ten-fold cross validation is used to determine good C and gamma parameter values, followed by actual classification using these values. The recognition rates achieved are in the range of 80-88 % correct classification.

VII. CONCLUSION

The system using SVMs as a classification tool in Sunspots Groups has shown promising results. The classification system relies on the multi-class SVMs classification tool and modified Zurich classification .The system improvement depends on the segmentation phase images quality. Thus high quality images and with less distortion will enhance the classification rate. The segmentation phase output is another important factor that has an effect on the classification rate.

REFERENCES

[1] T. Nguyen, C. P. Willis, D. J. Paddon,"Learning sunspot classification," Fundamental Informaticae XX, vol.72, 2006, pp. 295-309.

[2] P. McIntosh," Solar Physics,", pp. 125-251, 1990.

[3] S. Zharkov, V. Zharkova, S. Ipson, A. Benkhalil,"Technique for automated recognition of sunspots on full disk solar,", EURASIP Journal on Applied Signal Processing, vol. 15 , pp. 2573–2584 2005.

[4] S. Majed, M.Abd and V.Zharkova," Automated detection of sunspots and sunspot groups in full-disk solar images," ,Accepted in CISSE09

[5] Christopher J.C., "A tutorial on support vector machines for pattern recognition", [Online] Available: http://aya.technion.ac.il/karniel/CMCC/SVM-tutorial.pdf.

[6] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multi-class support vector machines", IEEE Transactions on Neural Networks, Vol. 13,No.2, pp. 415-425,March 2.002.

[7] Eugene A. Borovikov, "An evaluation of support vector machines as a pattern recognition tool", University of Maryland at College Park, 2005., [Online]. Available:www.umiacs.umd.edu/users/yab/SVMForPatternRecognition/report.pdf,

[8] Kreethi, S. S. and Lin C. J., "Asymptotic behaviors of support vector machines with gaussian kernel", Neural Computation, Vol. 15 No.7, pp. 1667-1689, 2003.

[9] C-W. Hsu, C-C. Chang, and C-J. Lin, "A practical guide to support vector classification", [Online]. Available: www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[10] M. Abd and G. Paschos," Effective arabic character recognition using support vector machines," In Proc.2007, Innovations and Advanced techniques in Computer and Information Sciences and Engineering, Springer, ISBN 978-14020-6267-4, 2007, PP. 7-11.

[11]. Solar Heliospheric Observatory (SOHO), (10-9- 2009),[Online]. Available: http://sohowww.nascom.nasa.gov/

[12] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," Journal of Electronic Imaging. Vol.13 (1), pp.146–165, 2004.

A Novel Ballistics Imaging System for Firearm Identification

Dongguang Li
School of Computer and Security Science, Edith Cowan University, Australia
Email: d.li@ecu.edu.au

ABSTRACT

The efficiency of traditional ballistics imaging system is heavily dependent upon the expertise and experience of end-user. An intelligent ballistics imaging system is highly demanded to address this issues. This paper presents a novel ballistics imaging system which achieves traditional functions and many new features including line-scan image module, characteristics extraction module and intelligent image processing module. Benefit from these features, the system can identify firearm more efficiently and effectively.

Index Terms— Ballistics system, firearm identification, image processing

1. INTRODUCTION

The analysis of marks on the cartridge case and projectile of a fired bullet is very important to identify the firearm from which the bullet is discharged [1, 2]. When a firearm is loaded and fired, the characteristic marks are produced on the cartridge case and projectile. These marks are significant evidences for identifying the suspicious firearm. Over thirty different features within theses marks can be referred as a 'fingerprint' for ballistics recognition [3]. Since each firearm owns its unique toolmarks [4, 5], it is possible to identify both the type and model of a firearm and each individual weapon as effectively as human fingerprint identification.

Several commercial or testing ballistics imaging systems have developed by organizations or researchers for decades [6-9]. Integrated Ballistics Identification System (IBIS) [9] is an international ballistics imaging system which has been widely used in several countries. Another commercial system called "Bulletproof" was developed by a Canadian company named Walsh Automation, which can acquire and store images of projectiles and cartridge cases by searching its image database for particular striations on projectiles and cartridge cases. However, it is required to match the impressed markings or striations by end-user on the projectiles and cartridge cases. This kind of limitation makes the system very difficult to use. The firearm identification system called Fireball [8] was developed by Edith Cowan University in 1995, and it has been used by Australia Police Services for identifying, storing, and retrieving the images of projectiles and cartridge cases. The major limitations of Fireball are that the shape and position are turned manually and the precision of comparisons relies on the expertise and experience of the end-user. The major drawbacks of the existing systems are outlined as follows: Firstly, the identification of firearm strongly depends upon the end-used due to lack high level intelligent image processing techniques; secondly, the image contrast of cartridge cases and projectiles is not good due to failure to find a satisfactory illuminator; finally, the lack of online image processing, retrieving and identification does not meet the demand of end-user. Therefore, the existing systems require further improvement and development.

The automatic ballistics imaging system mainly includes two components, i.e. the acquisition and the correlation component [10]. The acquisition component includes capturing of cartridge cases and projectiles and extracting their features in order to make them analyzable. The correlation components are responsible for comparing the features and organizing the results for the user's inspection. The following functions should be involved in ballistics imaging system [11]:

- Evaluate the degree of similarity between two sets of features;
- If more than two bullets are involved in a comparison, to organize the results of a set of comparisons in some convenient way;
- To provide the user with tools to verify the results obtained by the correlation algorithms.

The paper describes a novel ballistics imaging system for firearm identification. The system can intelligently capture, store, analyze and retrieve the digital images of cartridge cases and projectiles. The main contribution of the system is described as follows. Firstly, it improves efficiencies in time and personnel by computerizing imaging system; secondly, the precision of identification has been also improved due to the introduction of artificial intelligent techniques; finally,

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_57, © Springer Science+Business Media B.V. 2010

the online system is convenient for end-user and more independent on the client computer.

The rest of the paper is organized as follows: Section 2 gives an introduction to firearm identification. In Section 3, we illustrate the ballistics imaging system. An intelligent cartridge case segmentation algorithm is described in Section 4. Section 5 concludes this paper.

2. FIREARM IDENTIFICATION

Firearm identification aims to provide a link between the suspect firearm and the forensic ballistics specimens based on the depressions, scratches and markings on the specimens of cartridge cases and projectiles. These various marks are called 'toolmark' and it is thought that the toolmark is unique to itself [5]. In this regard, it is possible to identify the suspect firearm by toolmark left on ballistics specimens.

Firearm identification mainly includes the following stages [12]:

- The crime scene weapon is examined and test fired for a test specimen;
- The characteristic marks on the cartridge cases and projectiles are found and recorded;
- Features of the characteristic marks are investigated;
- Comparison of marks between the test specimen and the crime scene specimen;
- Assessment of the comparison between the test and crime scene specimens.
- Storage of the characteristic marks for further comparisons.

Table 1 Characteristics of a Firearm

Object	Class characteristics	Individual characteristics
Cartridge Case	Caliber Position of firing pin Position of extractor mark Position of ejector mark Breach or bolt impression	Striations of firing pin impression
Projectile	Caliber Twist direction of rifling Land mark width Groove mark width Degree of the rifling twist Number of lands Number of grooves	Striations of groove impressions

Charchman divided the characteristics of toolmarks on ballistics specimens into three major categories [13]:

- "C" type characteristics were defined as class characteristics;
- "A" type characteristics were defined as accidental characteristics, which are individual characteristics;
- "B" type characteristics were defined as broach series characteristics, with B1 and B2 subclasses that occur on the edges of land impressions.

Generally, Firearm identification is concerned with two types of characteristics, i.e. class characteristics (C type) and individual characteristics (A type). While broach series characteristic (B type) is only applied in particular types of rifles. Class characteristics are useful in determining what brand of firearm was used, but not helpful in identifying a individual firearm. Individual characteristics can directly match a bullet to a firearm. That is, the match between individual characteristics and a bullet can ensure that one certain firearm was the only firearm that could have fired the bullet. These characteristics are detailed in Table 1.

3. BALLISTICS IMAGING SYSTEM

Due to several drawbacks of workstation system including low data sharing and security, expensive cost of install and maintenance, inconvenient application environment etc, we developed an Online Ballistics Imaging System (OBIS) based upon Browse-Server (BS) structure. Fig. 1 shows the structure of the system. By a computer with internet access and a web browser, users can access OBIS anywhere and anytime.

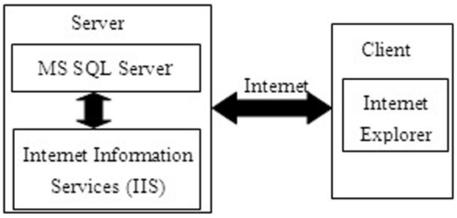


Fig. 1 System structure of OBIS

OBIS not only achieves the traditional functions including image capturing, storing, analyzing and retrieving but also provides intelligent image processing. The new functions of OBIS include image acquisition module, characteristic extraction module, firearm identification module and intelligent image processing module.

3.1. Image acquisition

The digital image acquisition of cartridge cases and projectiles focuses on the following issues: how to select and locate the satisfactory illuminator to get high contrast digital images; how to transform the surface of cylindrical projectile into a 2D flat image in order to facilitate identification.

As indicated by Smith [8], in order to produce high contrast of images, the illuminator must be installed at an incident angle of greater than 45 degrees from the illuminator to the plane of the head of the cartridge case. Illuminator from such angles can produce better shadow cast of the peaks and ridges of striations. The high quality of shadow cast makes discernible features for the identification process. According to Smith's guideline, we test point source illuminator and LED ring-light illuminator. Although point source illuminator can produce better shadow cast, the shadow cast of the image may vary with the rotational position of the cartridge case about its central axis. While the shadow cast of the image is invariant in LED ring-light illuminator. Therefore, LED ring-light illuminator is selected. Fig. 2 shows the high contrast image of head of cartridge case using LED ring-light illuminator.

Fig. 2 The image of head of cartridge case

The cylindrical shapes of ballistics projectiles (displayed in Fig. 3(a)) are inherently difficult to identify using traditional incident light microscope. Maintaining the quality of image from low magnification microscope is very difficult because of loss of focus as the ballistics projectile translated and rotated. To resolve the issue, the line-scan imaging technique [14] is employed by scanning consecutive columns of picture information and storing the data in a frame buffer so that a 2D image of the surface of the cylindrical projectile is produced (displayed in Fig. 3(b)). Benefit from the line-scan technique, all points on the 2D image are in focus, because the cylindrical projectile is rotated about an axis of rotation relative to a stationary line array of sensor. Thus, during one full rotation of the cylindrical ballistics projectile, all points on the rotating surface can be captured on the 2D image.

Fig. 4 shows the line-scan imaging system including stepper motor, LED ring-light illuminator, microscope, CCD camera and pc with graphic card. The procedure of line-scan imaging approach is detailed as follows:

1. With the stepper motor's every step;
2. The CCD camera captures the current image of projectile specimen and send the image to PC by graphic card;
3. The pixels on middle column in the image is extracted and saved consecutively in an array in the buffer on PC;
4. Repeat step 1, 2, and 3 until the whole surface of the projectile specimen is scanned.
5. Combine all arrays in the buffer one after another so that a 2D flat image for the whole surface of the projectile is produced.

(a) Image of projectile

(b) 2D image of projectile
Fig. 3 Image of ballistics projectile

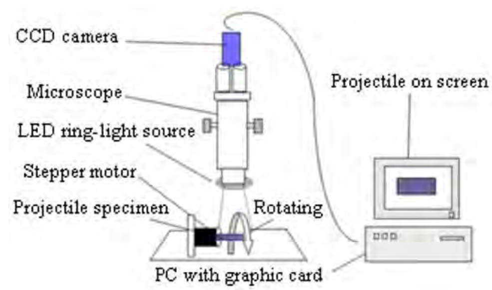


Fig. 4 Line-scan imaging system

3.2. Characteristic extraction

Since broach series characteristic (B type) is only applied in particular types of rifles, we focus on class characteristics and individual characteristics. Class characteristics on cartridge case and projectile are measured on corresponding images acquired by image acquisition module. Due to the lack of criteria of individual characteristics, we directly save the corresponding images into Database.

To facilitate measurement, we employ a clock face to describe class characteristics on cartridge case. Cartridge cases are divided into two classes, one is centrefire cartridge case, and the other is rimfire cartridge case. For rimfire cartridge case, the firing pin mark position is designated at 12 o'clock and the extractor and ejector mark positions are coded with respect to this position. While, for centrefire cartridge case the ejector and extractor mark position are orientated as they would be in the firearm. These characteristics can be easily measured and saved by mouse click on the graphic user interface.

It is difficult to accurately measure the class characteristics on projectile due to its cylindrical shape. On the contrary, these characteristics can be easily measured on 2D images acquired by line-scan imaging system. Fig. 5 shows that these characteristics can be measured by calculating the difference of coordinates of pixels. For example, "Land mark width" can be measured using the following steps: firstly, draw a horizontal line on "Land mark" in the image by mouse; then calculating the number of pixels between the starting pixel and the ending pixel of the line. Clearly, the product multiplying the number by the length of each pixel is the value of length of "Land mark width".

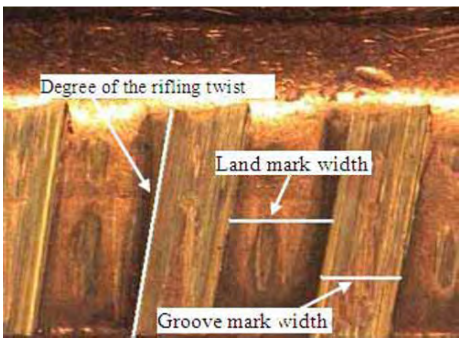


Fig. 5 Characteristics measurement on 2D image

3.3. Firearm Identification

In our system, firearm identification includes two steps: one is to retrieve the short list of suspect firearm by searching class characteristics saved in database. The other is to make sure the unique suspect firearm from the short list by comparing their high resolution images in detail manually.

The short list can be easily generated by graphics query interface. Various image techniques are provided to further compare the images in short list. Benefit from these image techniques, it is easy to find the suspect firearm. These techniques are listed as follows:

- Freely zoom, translate and rotate images;
- Transparentize and overlap images;
- Add or remove "image mask" including grid, clock face.

4. INTELLIGENT SEGMENTATION OF CARTRIDGE CASE

The segmentation of cartridge case from the input image (displayed in Fig. 6(a)) is of paramount importance for firearm identification. However, the satisfactory result cannot be achieved using traditional threshold-based methods because of the heavy shadow in the input image. To acquire the satisfactory segmented image of cartridge case, a novel approach based on color invariant properties [15] and geometrical shape of cartridge case is proposed. The method is composed of two steps. That is, remove the background and shadows using color invariant properties; detect and segment the circular cartridge case using Hough transform [16].

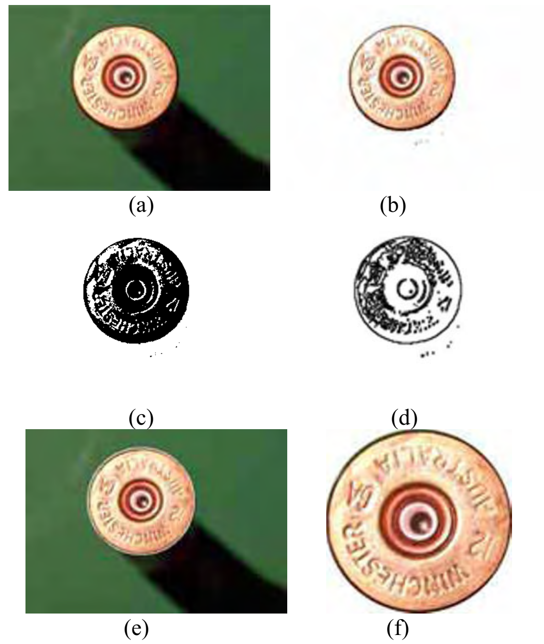


Fig. 6 Segmentation of cartridge case. (a) Input image, (b) Image removed background and shadows, (c) Binarized image, (d) Edged image,(e) Detected cartridge case (white circle), (f) Segmented cartridge case

4.1. Remove background and shadows

The color invariant properties are employed to remove the background and shadows. Color invariants are function which describes the color configuration of each pixel on images discounting shadows, and highlights. These features are demonstrated to be invariant to a change in the imaging conditions including viewing direction, object's surface orientation and illumination. Color invariants features are acquired by RGB using the following equations:

$$C1(x,y) = \arctan \frac{R(x,y)}{\max(G(x,y),B(x,y)}$$

$$C2(x,y) = \arctan \frac{G(x,y)}{\max(R(x,y),B(x,y)}$$

$$C3(x,y) = \arctan \frac{B(x,y)}{\max(R(x,y),G(x,y)}$$

where, R(x, y), G(x, y), and B(x, y) respectively represent the red, green and blue components of a pixel in a image, while the C1(x, y), C2(x, y) and C3(x, y) represent corresponding color invariants.

The procedure to remove background and shadows is detailed as follows:

1. Select a background whose values of color invariants do not overlap the values of cartridge case. It is well known that most cartridge cases are made by brass. Therefore, they have similar values of color invariants. In our research, the range of color invariants for cartridge cases is $C_1 \in [0.71,1.11], C_2 \in [0.4,0.80], C_3 \in [0,0.80]$, the color with R=54, G=12, B=170 is selected as background. Its color invariants is C1= 0.31, C2=0.07, C3=1.26.
2. Capture the image of cartridge case based on the background selected in step 1.
3. Remove the pixels whose color invariants near to the values of background. Thus, the background and shadows are removed (Fig. 6(b) shows the result).

4.2. Segment cartridge case

Although the most background and shadows are removed in the Section 4.1, the image of cartridge case isn’t suitable for identification. A few noise pixels belong to the background does not be removed, and the borders of the cartridge case are not smooth. To address these issues and accurately segment cartridge case, the Hough transform is employed. The method segmenting cartridge case includes the following steps:

1. Binarize the image displayed in Fig. 6(b) using Otsu algorithm (shown in Fig. 6(c));
2. Detect the edges to the binarized image using Sobel operator (shown in Fig. 6(d));
3. Detect the circle center and radius of the cartridge case in the edged image using Hough transform (shown in Fig. 6(e));
4. Segment the cartridge case based the circle center and radius from the input image (shown in Fig. 6(f)).

5. CONCLUSION

This paper presents a novel ballistics imaging system for firearm identification which not only involves traditional functions including image capturing, storing, analyzing and retrieving but also intelligent image processing. The system can identify firearm more efficiently and effectively due to the new functions including the line-scan imaging system and intelligent image processing. Future work includes researching the criterion for extracting individual characteristics and increasing more intelligent analysis for firearm identification.

Acknowledgement

The ballistics specimens were provided by Australia Police. Mr Z. Huang , as a visiting fellow at ECU, helped with digital image acquisition.

REFERENCES

[1] Dongguang Li, "Ballistics Projectile Image Analysis for Firearm Identification," *IEEE Transaction on Image Processing*, vol. 15, pp. 2857-2865, 2006.

[2] R. Saferstein (ED), "Forensic Science Handbook," *Volume 2. Englewood Cliffs: Prentice Hall*, 1988.

[3] G.Burrard, "Identification of Firearms and Forensic Ballistics," *London, U.K:Herbert Jenkins*, 1951.

[4] A. Biasotti and J. Murdock, "Criteria for identification or State of the Art of Firearm and Toolmark Identification," *AFTE Journal*, vol. 16, pp. 16-34, 1984.

[5] A. Schwartz, "A Systemic Challenge to the Reliability and Admissibility of Firearms and Toolmark Identification," *The Columbia Science and Technology Law Review*, vol. VI, pp. 1-42, 2005.

[6] Chenyuan Kou, Cheng-Tan Tung, et al., "FISOFM: Firearms Identification based on SOFM model of Neural Network," in *Security Technology, 1994. Proceedings. Institute of Electrical and Electronics Engineers 28th Annual 1994 International Carnahan Conference*, Albuquerque, NM, USA, 1994, pp. 120-125.

[7] C.L. Smith, Robinson. M, et al., "Linescan Imaging for the positive identification of bal-listics Specimens," in IEEE *34th Annual 2000 International Carnahan Conference*, Ottawa, Ont., Canada, 2000, pp. 269-275.

[8] C. L. Smith, "Fireball: a forensic ballistics imaging system," in T*he Institute of Electrical and Electronics Engineers 31st Annual 1997 International Carnahan Conference*, Canberra, 1997, pp. 64-70.

[9] IBIS (Integrated Ballistics Identification System). http://www.fti-ibis.com/.

[10] Natalie G. Carpenter, "An Examination Of A Three-Dimensional Automated Firearms Evi-dence Comparison," in *Department of Criminology College of Arts and Sciences* University of South Florida, April 8, 2004.

[11] Bachrach B., "Development of a 3D-based automated firearms evidence comparison system," *Journal of Forensic Sciences*, vol. 47, pp. 1-12, 2002.

[12] C. L. Smith, "Multi-dimensional cluster analysis of class characteristics for ballistics specimen identification," in *Security Technology, 2001 IEEE 35th International Carnahan Conference*, London, UK, 2001, pp. 115-121.

[13] J. Churchman, "The reproduction of characteristics in signatures of Cooey rifles," *RCMP Gazette*, vol. 11, pp. 133-140, 1949.

[14] A. Zographos, M. Robinson, et al., "Ballistics Identification Using Line-Scan Imaging Tech-niques, in *Security Technology, 1997. Proceedings. The Institute of Electrical and Electronics Engineers 31st Annual 1997 International Carnahan Conference*, Canberra, ACT, Australia, " 1997, pp. 82-87.

[15] T. Gevers, "Color-based object recognition," *Pattern Recognition*, vol. 32, pp. 453-464, 1999.

[16] R. O. Duda and P.E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," *Comm. ACM*, vol. 15, pp. 11-15, 1972.

Color Seal Segmentation and Identification

Dongguang Li

School of Computer and Security Science, Edith Cowan University, Australia
E-mail: d.li@ecu.edu.au

***Abstract*—Automatic seal imprint identification system is highly demanded in oriental countries. Even though several seal identification techniques have been proposed, it is seldom to find the papers on the recovery of lost seal imprint strokes caused by superimposition. Most identification methods are conducted by matching the reference seal imprint and candidate one pixel by pixel. In this paper, a robust color seal segmentation and identification approach is proposed. This approach segments the seal imprint from the input image in terms of the adaptive thresholds. The lost seal imprint strokes are recovered based on the text stroke width that can be detected automatically. In addition, the moment-based seal identification is to compare the reference seal imprint and the recovered one. Experimental results show that the proposed method is able to correctly segment seal imprint and efficiently discriminate the genuine and forgery seal imprint.**

Keywords-Seal segmentation, Seal identification, Image binarization, Moment invariant

I. INTRODUCTION

Seals are widely used in oriental countries such as Chinese, Japanese and Korea to identify a person, a social group, or an organization etc[1, 2]. Currently, a variety of seals must be manually registered, which relies on the expertise and experience of professional technicians. The process of seal identification not only consumes so much time, but also inevitably introduces some misidentification. Therefore, a seal imprint verification system is highly demanded to resolve this issue automatically, speedily and reliably.

Seals can be made of various materials e.g. wood, stone, marble, plastics, ivory, crystal, metal. The shape and size of seals are also different. Basically, there are three shapes, i.e., circular, rectangular, and elliptic[3]. Generally, the color of seal imprint is pure red or pure blue. Seal identification can be regarded as a template match between the reference seal imprint and the candidate one. To exactly match the seals, a variety of features have to be considered, including the characters in the imprint, the scales, distributions, relative spatial positions, strokes etc. As a result, the matching is called "hard-matching" [4]. Currently, hard-matching is usually performed by human visual inspection. To describe the process of seal identification in detail, we define two variables R (Reference seal imprint) and C (Candidate seal imprint). To identify whether C is genuine or forgery, a professional technician firstly overlap the centroid of R and C, then rotate C to make it have the same orientation as R, and compare the corresponding details between them. If all of the details are satisfactory to the technician's subjective judgment, the candidate seal imprint C is regarded as a genuine one.

According to the process mentioned above, the hard-matching consists of two steps, i.e. segment seal imprint from the input image, and then identify whether the segmented one is genuine or forgery. There are three main issues on computerizing this procedure. Firstly, the input seal imprint is usually superimposed by other information such as signature or background text, consequently, the segmentation of seal imprint is difficulty[5]. Secondly, the difference between the genuine seal imprint and the forgery one is small, so the results are very sensitive to noise. Thirdly, it is very difficult to make the candidate seal imprint have exact same orientation and size as the reference one.

Automatic seal segmentation have been studied in many literatures [6-9]. These methods were mostly implemented based upon RGB or HSV color space and can acquire satisfactory results. However, none of the methods focuses on how to detect and recover losses of seal imprint strokes in segmented seals. The losses of seal imprint strokes often exist in segmented seal imprint due to superimposition among the seal, signature and background text. The result of seal identification may be strongly affected by losses of seal imprint strokes. As a consequence, it is of great importance in automatic seal imprint identification.

Several seal identification methods have also been proposed for decades [10-14]. These methods mainly include four steps: 1) Rotate the candidate seal imprint to the same orientation as the reference one; 2) Compute the centroid to the candidate imprint and reference one; 3) Overlap the rotated candidate seal imprint and the reference one in terms of the centroid; 4) Match them pixel by pixel. From these steps, we can see that the performance and robustness of seal identification strongly depend upon the precision to detect the angle of rotation and position of centroid. Owing to the difficulty in exactly overlapping the input seal imprint and reference one, the method matching candidate seal imprint and reference one pixel by pixel is difficult to use. Clearly, the exact detection of rotated angle is very complicated and difficult.

In this paper, a robust approach for seal imprint segmentation and identification is proposed. To improve the precision and robustness to segment seal imprint, an adap-

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_58,

tive thresholding method based on RGB components is presented. This segmentation method is very suitable for both seal segmentation and extracting a particular color cluster from an image. The recoveries for the losses of seal imprint strokes are considered for minimizing the errors in seal imprint identification. A moment-based method is employed because it is independent on the image translation, rotation and scaling. As a result, the proposed method not only avoids the complicated match processing but also improves the robustness and precision to identify seal imprints. The flowchart of the proposed method is shown in Fig. 1. It consists of five major steps: 1) Segment the seal imprint in terms of the adaptive thresholds; 2) Recover the losses of seal imprint strokes in segmented seal imprint; 3) Binarize the segmented seal imprint and the reference one using Otsu algorithm[15]; 4) Compute the Hu's moment invariants[16] to the segmented seal imprint and the reference one; 5) Identify them based on Euclidean distance of moment invariants.

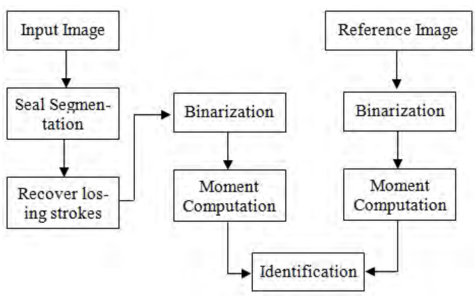


Fig. 1. Seal imprint identification system

The rest of the paper is organized as follows: Section 2 gives an introduction to related works. In Section 3, we illustrate the proposed approach. The experimental results are presented and analyzed in Section 4. Section 5 concludes this paper.

II. RELATED WORK

A variety of seal identification techniques have been proposed. In general, the studies mainly focus on the seal segmentation and seal registration. Seal segmentation deals with how to exactly extract seal imprint from the input image such as bank check image. Seal registration is to identify that the input seal imprint is genuine or forgery by comparing the input and the reference seal imprint.

Katsuhiko[9] extracted the signature and seal imprint by projecting orthogonally all pixels of a bank check image on to an appropriate axis in RGB color space[17]. This method may fail if the center of three clusters is all in or near a straight line. In [14], the input image is first transformed into HSV (Hue, Saturation, and Intensity-Value) perceptual space[18], and then K-means algorithm[19] is employed to divide the input image into three clusters corresponding to seal imprint, signature and background. Seal imprint is extracted based on the cluster index. Since K-means method divides the pixels in terms of Euclidean distance, the completely different color may be divided into a same cluster. For example, the color (R=255, G=0, B=0) and (R=0, G=255, B=0) are red and green, but they have the same Euclidean distance to black (R=0, G=0, G=0). This method may also get incorrect result when noise is strong. Liang[6] divided the RGB space into eight subspaces corresponding to color index 0, 1, 2 ,…, 7 by dichotomy, and then compute the color index for each pixels in input image in terms of equation (2.1), where "[]" represents integer Round operator, r, g, b represent respectively red, green and blue components in color normalized image. Finally, extract the seal imprint by index (e.g. the index of red is 1). This method may also be easily attacked by noise.

$$\text{Index}= [b/128] \times 4 + [g/128] \times 2 +[r/128] \quad (2.1)$$

Shih-Hsu [12] proposed a register method in terms of the feature points. This method includes four major steps: 1) Feature points extraction; 2) Estimate the rotation and translation parameters using the point matching technique; 3) Seal registration using the affine transformation by the estimated parameters; and 4) Fake detection from the registered seals. However, the seal imprints with different orientation stamped by the same seal may generate different feature points that may strongly affect the stability of this algorithm. To rotate the input seal imprint and the reference one to the same orientation, Yung-Sheng[10] used the contour analysis to find the principal orientation of a seal imprint, and then rotate the input seal imprint to the same orientation with the reference one. The algorithm may misidentify if the seal imprint does not exist principal orientation such as the seal imprint displayed in Fig. 2(b).

In this paper, we employ the Otsu algorithm [15] for the adaptive seal segmentation. Otsu is a typical and popular image binary algorithm based on global threshold. This method divides input data into two classes by threshold. The basic idea of Otsu is to find an optimal threshold that gives the best separation between the two classes in terms of their values. The threshold can be estimated automatically on an iterative scheme, in order to find the maximum variance between the two classes. It works well in situations where there is a reasonably clear valley between the modes of the histogram related to the two classes.

Hu's moment invariants[16] are very effective on identifying the input seal imprint. Moment invariants have been widely applied to image pattern recognition in a variety of applications due to its invariant features on image translation, scaling and rotation. Two-dimensional (p+q)th order moments are defined in equation (2.2).

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)dxdy \qquad p,q = 0,1,2,\cdots \quad (2.2)$$

To achieve the independence for image translation and rotation, the equation (2.2) is transformed into central moment shown in equation (2.3)

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\bar{x})^p(y-\bar{y})^q f(x,y)dxdy \qquad (2.3)$$
$$p,q = 0,1,2,\cdots$$

where $\bar{x}=m_{10}/m_{00}$ and $\bar{y}=m_{01}/m_{00}$, which are the centroid of the image. Scale invariance can be obtained by normalization. The normalized central moments are defined in equation (2.4)

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}, \quad \gamma = (p+q+2)/2, \ \mathrm{p+q} = 2,3,\cdots \quad (2.4)$$

Based on normalized central moments, Hu[16] introduced seven moment invariants:

$$\varphi_1 = \eta_{20} - \eta_{02}$$
$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$
$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \mu_{03})^2$$
$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \mu_{03})^2$$
$$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$
$$+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$
$$\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$
$$+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$
$$\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$
$$- (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

This seven moment invariants are useful properties of being unchanged under image scaling, translation and rotation.

III. PROPOSED APPROACH

In general, the shapes of seal may be circular, rectangular, or elliptic. The colors of seal imprint may be red or blue. For simplicity, only circular seal and red seal imprint are selected in our demonstration. The genuine seal imprint is shown in Fig. 2.(a). Obviously, the image is composed of three color clusters, i.e. the red seal imprint, the black text and the white background. Fig. 2.(b) displays the reference seal imprint in which only includes white background and red seal imprint. The forgery seal imprint is displayed in Fig. 2.(c).

(a) Genuine seal imprint (b) Reference seal imprint (c) Forgery seal imprint

Fig. 2. The input images for demonstration

A. Seal Segmentation

Seal segmentation aims to extract the seal imprint from the input image and recover the losses of seal imprint strokes caused by superimposition. It mainly contains the following steps: 1) Estimate the thresholds to segment the seal imprint, 2) Segment the seal imprint and text from the input image, 3) Recover the losses of seal imprint strokes, 4) Binarize the segmented seal imprint and reference one.

1. Estimate Thresholds

In our demonstration, the color of seal imprint is red, i.e. the red component of seal imprint is stronger than green and blue components. For three color clusters, i.e. red (seal imprint), white (background) and black (text), the R (Red), G (Green) and B (Blue) components are equal for white(R=G=B=255) and black(R=G=B=0), The red seal imprint can be easily segmented under the conditions R>G and R>B. Fig. 3 shows the segmented result only based on the conditions R>G and R>B. The result shows that so many pixels belonging to background or text are also meet the conditions in practice because of various noisy factors such as camera, illumination.

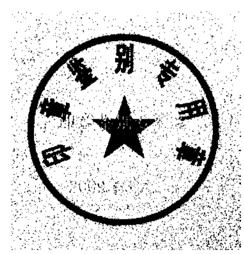


Fig. 3. Segmented Seal with Noise

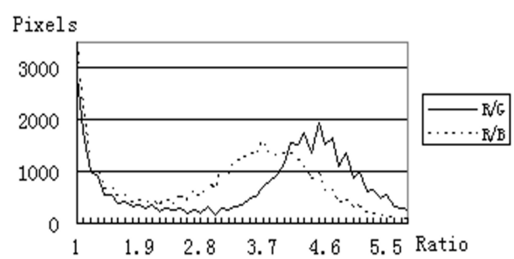


Fig. 4. Histograms for the values of R/G and R/B

We can found that the ratios of R/G and R/B for noise pixels are far less than the ratios of seal imprint pixels. It is possible to find the thresholds of R/G and R/B to divide the pixels which meet the conditions (R>G and R>B) into two classes, i.e. the noise pixels and the seal imprint pixels. To find the optimal thresholds for R/G and R/B, the histograms for the ratios of R/G and R/B on pixels whose R>G and R>B are constructed, as shown in Fig. 4. Since both histograms of R/G and R/B have evident bimodality, the optimal thresholds can be acquired using Otsu algorithm [15] detailed in Section 2. For Fig. 2(a), the optimal thresholds of R/G and R/B calculated by Otsu are respectively 2.1 and 1.8.

2. Segment Seal Imprint and Text

After acquiring the thresholds, the seal imprint can be segmented easily. Since the text in input image will be used in subsequent process, the seal imprint and text are segmented in the same time. The steps of segmentation are detailed in algorithm 1.

Algorithm 1 Segmenting Seal Imprint and Text from the Input Image

Input: an image with red seal imprint, black text and white background

Output: an image of red seal imprint, an image of black text

i: Estimate the threshold TH_{RG} and TH_{RB} using method described above;

ii: Scan input image, get values of the red, green and blue component in each pixel and save in variable r, g and b respectively;

iii: if $r/g > TH_{RG}$ and $r/b > TH_{RB}$

The pixel is belong to seal imprint I_{seal}

Else

The pixel is belong to background or text I_{BT}

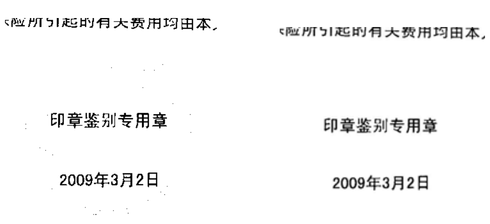
End if;

iv: Binarize the pixels in I_{BT} using Otsu algorithm [15] and then get the text image I_{text};

v: Remove the black block in I_{text} whose size less than 5 pixels;

vi: Return the seal imprint I_{seal} and black text I_{text}.

The segmented seal imprint is displayed in Fig. 7(a). And the black text with noise is shown in Fig. 5(a). To remove the noise, the size of each black block is computed. And then remove the small block whose size less than 5 pixels. After this operation, a clear text image is acquired and showed in Fig. 5(b).



(a) Text with noise (b) Noise-removed text

Fig. 5. Images of black text

3. Recover the losses of seal imprint strokes

Since the brightness and color vary in superimposition of text and seal imprint inks, the character strokes of seal imprint in superimposition may lose in segmented seal imprint. This may cause incorrect result in seal imprint registration process. To avoid this, a method for detecting and recovering losses of seal imprint strokes is proposed in this section. Comparing the Fig. 7(a) and Fig. 5(b), we can see that the losses of seal imprint strokes basically take place in superimposition between the seal imprint and text. Furthermore, the width of lost stokes is very close to the width of stokes corresponding text. Accordingly, we apply the following steps to recover the lost strokes of seal imprint.

- Estimate the width of the strokes of black text;
- Recover the lost strokes in seal imprint image in terms of the width of the strokes of black text.

To estimate the stroke width of black text, we scan each row of black text image (shown in Fig. 5(b)) to calculate the number of each continuous black pixels, and then construct the stroke histogram using all of the number of continuous black pixels. The position of maximum in histogram is considered as the stoke width of black text. For the image shown in Fig. 5(b), the histogram is displayed in Fig. 6. Evidently, the stroke width of black text is 4 pixels.

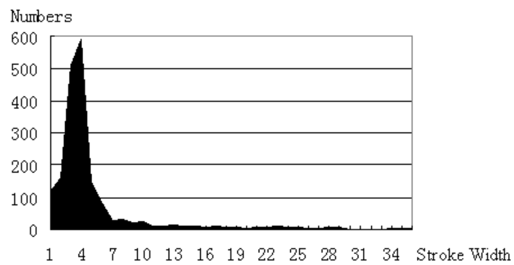
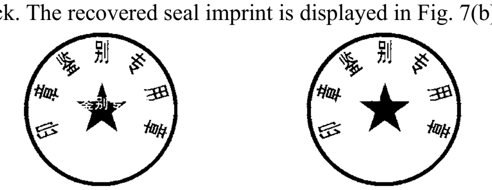


Fig. 6. Histogram of stroke width on black text

After getting the stroke width of black text W_{stroke}, we can recover the segmented seal imprint using the following procedures. Scan the black text image shown in Fig. 5(b). For each black pixel, detect its $W_{stroke} \times W_{stroke}$ neighbor pixels, if the pixels located at up and down position or left and right position of the black pixel are all red, the white pixel with the same position as the black text image in the seal imprint image (shown in Fig. 7(a)) is converted into black. The recovered seal imprint is displayed in Fig. 7(b).

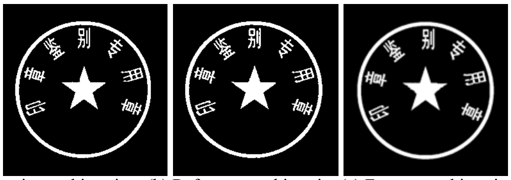


(a) Seal imprint without strokes recovery (b) Recovered seal imprint

Fig. 7 Segmented seal imprint

B. *Seal Registration*

Normally, the registration includes two steps: 1) Accord the orientation and size of candidate seal imprint with the reference one; 2) Overlap the candidate seal imprint and the reference one in terms of their centroid, and then match them pixel by pixel. Since exact adjustment on the orientation of candidate seal imprint is very difficult or even impossible, especially to circular seal, the precision of registration may be strongly affected by adjusted precision. We propose a novel moment-based approach to address the difficulty. The moment invariants are independent of translation, rotation and scaling, so the complicated adjustment in orientation and size can be avoided.



(a) Genuine seal imprint (b) Reference seal imprint (c) Forgery seal imprint

Fig. 8. Binarized seal imprint

To minimize the effect of luminance, we transform all seal imprints into binary images with black background and white seal imprint. The sample seal imprints are displayed in Fig. 8. Seal registration is achieved by computing the seven moment invariants on all seal imprints and Euclidean distance (shown in equation (3.1)) based on these invariants.

$$D = \sum_{i=1}^{7} \left| \frac{C_i - R_i}{R_i} \right| \qquad (3.1)$$

where C_i, R_i respectively present the seven moment invariants of candidate seal imprint and reference one. Table 1 displays the results of comparison between the reference seal imprint and genuine or forgery one. Although it is very difficult to manually distinguish the reference seal imprint from the forgery one (displayed in Fig. 8), the Euclidean distance between the reference seal imprint and forgery one is evidently greater than the values between the reference seal imprint and genuine one. Consequently, the forgery seal imprint can be easily identified by the Euclidean distance.

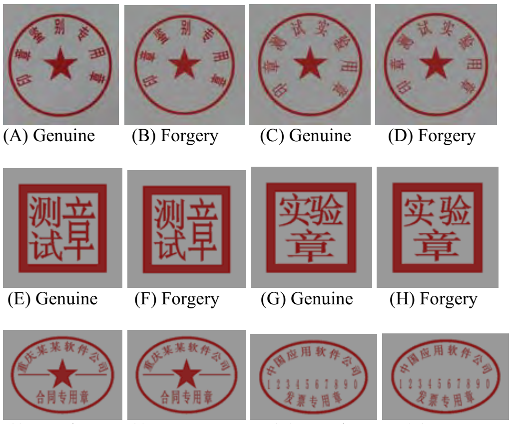
Table 1. Comparison of moment invariants for reference, genuine and forgery seal imprint

seal	Moment invariant							Distance to reference
Reference	0.00385	8.93E-8	7.94E-11	5.52E-10	-1.15E-19	-1.52E-13	-1.14E-20	0
Genuine	0.00388	8.98E-8	8.1E-11	5.64E-10	-1.19E-19	-1.54E-13	-1.51E-20	0.427
Forgery	0.00379	5.22E-8	8.94E-11	4.11E-10	-0.78E-19	-0.89E-13	-1.01E-20	1.663

IV. Experimental Results

To evaluate the performance of the proposed seal segmentation and identification algorithm, extensive experiments are conducted based on sample images. Twelve seals are used in the experiments (displayed in Fig. 9). Seals A to D are circular seals (Type 1), Seals E to H are rectangular seals (Type 2), and Seal I to L are elliptic seals (Type 3). Each type includes two genuine seals and two forgery seals. All seals are engraved by one maker, therefore the difference between the genuine seal and forgery one are very small. The experimental steps are describes as follows.

- Stamp each genuine seal on white paper, and convert it into a digital image by CANON IXUS 90 IS camera as a reference seal imprint. Then transform them into binary images and calculate the moment invariants on the binary images;
- Arbitrarily stamp each of twelve seals on five different documents. And then convert them into digital images by CANON IXUS 90 IS camera. Thus, sixty candidate seal imprints are applied to examine the proposed algorithm;
- Segment each seal imprint from input images with seal imprint and recover losses of seal imprint strokes. Then, transform them into binary images;
- Calculate the moment invariants on each image from last step.
- Calculate the Euclidean distance between the reference seal imprint and corresponding candidate seal imprints.
- Make sure whether the candidate seal imprint is genuine or not in terms of the selected threshold.



(A) Genuine (B) Forgery (C) Genuine (D) Forgery

(E) Genuine (F) Forgery (G) Genuine (H) Forgery

(I) Genuine (J) Forgery (K) Genuine (L) Forgery

Fig. 9. Twelve seals for experiments

Fig. 10 shows the results of recovered seal imprints segmented from the input image (only Seal A and B are displayed). Although the orientation of seal imprints are different from each other, which not affects the results of identification due to the moment invariants are independent to rotation. The experimental results are displayed in Table 2. From this table, we can see that the Euclidean distance between genuine and reference seal imprints are all less than 1, and the Euclidean distance between the reference and forgery seal imprints are all greater than 1.8. In real application, the errors to identify a forgery seal imprint as a genuine one is not acceptant. Therefore, in our experience, we select 1 as the threshold to identify whether the seal imprint is genuine or forgery.

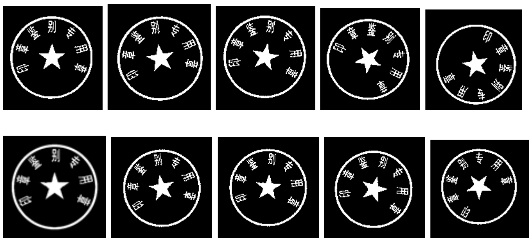


Fig. 10. Five seal imprints of A (row 1) and five seal imprints of B (row 2)

Table 2. Comparison of moment invariants for twelve seals

Groups	Distance for genuine seals					Distance for forgery seals				
	1	2	3	4	5	1	2	3	4	5
A and B	0.436	0.398	0.346	0.480	0.289	2.21	2.25	2.26	2.12	2.14
C and D	0.275	0.710	0.699	0.837	0.418	11.95	11.61	12.75	13.33	11.72
E and F	0.904	0.277	0.185	0.412	0.125	1.87	2.24	2.12	2.24	2.65
G and H	0.100	0.100	0.160	0.332	0.211	3.02	3.13	3.02	3.02	2.97
I and J	0.328	0.593	0.408	0.384	0.758	2.12	2.12	2.08	2.16	2.51
K and L	0.449	0.588	0.395	0.340	0.451	2.10	2.41	2.01	2.15	1.98

Our Proposed approach has been implemented using Visual Basic .NET, and running on a PC (CPU P4 2.0G, RAM 1G). The processing time to segment and recover the seal imprint is about 0.9 second for 500*500 color image. And the processing speed for calculating the moment invariants is about 0.8 second. As a result, the whole computational time is less than 2 seconds.

V. Conclusion

In this paper, a robust segmentation and identification algorithm for color seal imprint is proposed. This method is composed of segmentation of seal imprint, recovery of losses of seal imprint strokes, and identification of seal imprint. The segmentation automatically extract seal imprint from the color input image in terms of the adaptive thresholds. To minimize the errors, the losses of seal imprint strokes are recovered by detecting the width of black text stroke and superimposition pixels between the red seal imprint strokes and the black text strokes. To avoid complicated adjustment on orientation and size between the reference seal imprint and candidate one, the moment-based technique is applied to seal imprint identification.

Experimental results showed that the proposed segmentation method is able to segment seal imprint from color input image, and the losses of seal imprint strokes can be correctly recovered. Moreover, the identification method has high performance in exactly discriminating the genuine and forgery seal imprint. The computational time is less than 2 seconds on general PC. Therefore, the proposed approach is fast and feasible. In future, we will find more features on seal imprints, and employ neural network instead of Euclidean distance to identify seal imprints. In addition, more experiments will be conducted to check whether the approach is able to be applied to more extensive conditions.

VI. Acknowledgement

Mr Z. Huang, as a visiting fellow at ECU, helped with digital image preparation.

References

[1] H. Leung, "Analysis of traditional Chinese seals and synthesis of personalized seals," in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, 2004, pp. 1283-1286 Vol. 2.

[2] J. J. Brault and R. Plamondon, "Segmentation handwritten signatures at their perceptually important points," *IEEE Trans. Pattern. Anal. Mach. Intell.,* vol. 15, pp. 953-957, 1993.

[3] Chou Ju-Pei and Chen Hu-Sheng, "Chinese seal identification," in *Security Technology, 1990. Crime Countermeasures, Proceedings. IEEE 1990 International Carnahan Conference on*, 1990, pp. 51-53.

[4] Yung-Sheng Chen, "Computer Processing on the Identification of a Chinese Seal Image," in *Third International Conference on Document Analysis and Recognition (ICDAR'95)*, 1995, pp. 422-425.

[5] Yung-Sheng Chen, "Approach to the Chinese Seal Registration," in *Advances in Communication Systems and Electrical Engineering*. vol. 4, 2008, pp. 529-542.

[6] Liang Cai and Li Mei, "A Robust Registration and Detection Method for Color Seal Verification," in *Advances in Intelligent Computing*. vol. 3644, 2005, pp. 97-106.

[7] Ching-Liang Su, "Chinese-Seal-Print Recognition by Color Image Dilating, Extraction, and Gray Scale Image Geometry Comparison," *Journal of Intelligent and Robotic Systems* vol. 44, pp. 349-359, 2005.

[8] Aureli Soria-Frisch, "Color Seal Extraction from Documents: Robustness through Soft Data Fusion," *EURASIP Journal on Applied Signal Processing,* vol. 2005:13, pp. 2146-2152, 2005.

[9] K. Ueda, "Extraction of signature and seal imprint from bankchecks by using color information," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, pp. 665-668 vol. 2.

[10] Yung-Sheng Chen, "Registration of Seal Images Using Contour Analysis " in *Lecture Notes in Computer Science*. vol. 2749: Springer Berlin / Heidelberg, 2003, pp. 255-261.

[11] Ching-Liang Su, "Edge distance and gray level extraction and orientation invariant transform for Chinese seal recognition," *Applied Mathematics and Computation,* vol. 193, pp. 325-334, 2007.

[12] Shih-Hsu Chang, Fang-Hsuan Cheng, et al., "Automatic seal identification by using point matching and fake detection," *Pattern Recognition Letters,* vol. 20, pp. 1545-1552, 1999.

[13] T. Horiuchi, "Automatic seal verification by evaluating positive cost," in *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, 2001, pp. 572-576.

[14] K. Ueda, T. Mutoh, et al., "Automatic verification system for seal imprints on Japanese bankchecks," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, 1998, pp. 629-632 vol. 1.

[15] N. Otsu, "Threshold Selection Method from Gray-Level Histograms," *IEEE Transcations on Systems Man and Cybernetics,* vol. 9, pp. 62-66, 1979.

[16] Hu Ming-Kuei, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions,* vol. 8, pp. 179-187, 1962.

[17] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing (Third Edition)*: Prentice Hall, 2007.

[18] A R. Smith, "Color gamut Transformation Pairs," *Computers Graphics,* vol. 12, pp. 12-19, 1978.

[19] Y. Linde, A.Buzo, et al., "An algorithm for vector quantizer design," *IEEE Trans. on Communications,* vol. 28, pp. 84-95, 1980.

Efficient SOPC-based Multicore System Design using NOC

Vignesh Veerapandian, Xingguo Xiong
Department of Electrical and Computer Engineering, University of Bridgeport, Bridgeport, CT 06604, USA
Email: vveerapa@bridgeport.edu, xxiong@bridgeport.edu

***Abstract* — Due to the advancement of VLSI (Very Large Scale Integrated Circuits) technologies, we can put more cores on a chip, resulting in the emergence of a multicore embedded system. This also brings great challenges to the traditional parallel processing as to how we can improve the performance of the system with increased number of cores. In this paper, we meet the new challenges using a novel approach. Specifically, we propose a SOPC (System on a Programmable Chip) design based on multicore embedded system. Under our proposed scheme, in addition to conventional processor cores, we introduce dynamically reconfigurable accelerator cores to boost the performance of the system. We have built the prototype of the system using FPGAs (Field-Programmable Gate Arrays). Simulation results demonstrate significant system efficiency of the proposed system in terms of computation and power consumption. Our approach is to develop a highly flexible and scalable network design that easily accommodates the various needs. This paper presents the design of our NOC (Network on Chip) which is a part of the platform that we are developing for a reconfigurable system. The major drawback of SOPC based systems lies in the routing of the various on-chip cores. Since it is technically difficult to integrate more than one core on a single chip, we come across several routing problems which lead to inefficient functioning. Thus we implemented several NOC based routing algorithms which considerably improve accessing speed and enhance the system efficiency.**

Keywords: Multicore system, System on a Programmable Chip (SOPC), Network on Chip (NOC), Multiprocessor System-on-Chip (MPSOC).

I. INTRODUCTION

During the 1990s more and more processor cores and large reusable components have been integrated on a single silicon die, which has become known under the label of System-on-Chip (SOC). Buses and point-to-point connections were the main means to connect the components, Hence they can be used very cost efficiently. As silicon technology advances further, several problems related to buses have appeared. [1][5] Buses can efficiently connect 3-10 communication partners but they do not scale to higher numbers. As a result, around 1999 several research groups have started to investigate systematic approaches to the design of the communication part of SOCs. It soon turned out that the

Problem has to be addressed at all levels from the physical to the architectural to the operating system and application level. Hence, the term Network on Chip (NOC) is today used mostly in a very broad meaning, encompassing the hardware communication infra-structure, the middleware and operating system communication services and a design methodology and tools to map applications onto a NOC. All this together can be called a NOC platform. [4] Networks on Chip (NOCs) have emerged as a viable option for designing scalable communication architectures. For multiprocessor System-on-Chips (MPSOCs), on-chip micro networks are used to interconnect the various cores. The main idea with NOCs, besides the solutions to the physical issues, is the possibility for more cores to communicate simultaneously, leading to larger on-chip bandwidths. The adoption of NOC architecture is driven by several forces: from a physical design viewpoint, in nanometer CMOS technology interconnects dominate both performance and dynamic power dissipation, as signal propagation in wires across the chip requires 2 multiple clock cycles. NOC links can reduce the complexity of designing wires for predictable speed, power, noise, reliability, etc., thanks to their regular, well controlled structure. From a system design viewpoint, with the advent of multi-core processor systems, a network is a natural architectural choice.

NOC can provide separation between computation and communication; support modularity and IP reuse via standard interfaces; handle Synchronization issues; serve as a platform for system test and hence increase engineering productivity.

II. DIFFERENT NOC TOPOLOGIES

The Network-on-Chip (NOC) architecture, as outlined in Figure 1, provides the communication infrastructure for the resources. In this way it is possible to develop the hardware of resources independently as stand-alone blocks and create the NOC by connecting the blocks as elements in the network.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_59,

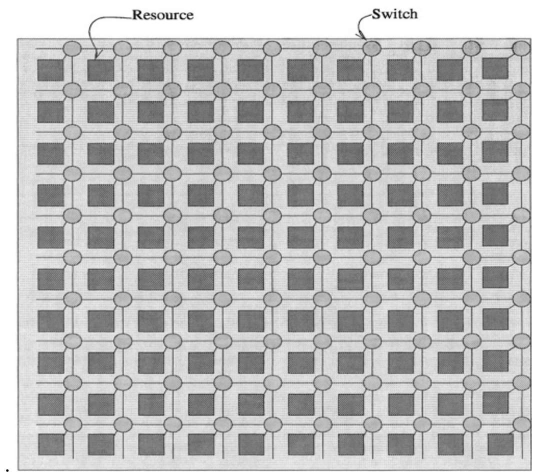


Figure.1. Network on chip [7]

A number of different NOC topologies have been proposed. They all have in common that they connect resources to each other through networks and that information is sent as packets over the networks [7]. Network on Chip (NOC) has evolved as an important research topic during the last few years. The idea is that scalable switched networks are used for on-chip communication among processing units, in order to cope with design of continuously growing systems. Design complexity promotes reuse of investment in earlier designs as well as purchase of outside intellectual property (IP). However, in larger designs, communication among components will become a bottleneck using traditional techniques like common buses. NOC is one solution to address this issue because packet switched communication can provide higher flexibility, throughput and reusability. To gain full advantage when using this concept in NOC architecture design, the size of resources should be similar and the communication facilities should be homogeneous.

2.1 Honey Comb Technology

In NOC design, the resources communicate with each other by sending addressed packets of data and routing them to the destinations by the network of switches [7]. Though many topologies are possible, we will first discuss about Honey comb topology. The overall organization is in the form of a honeycomb, as shown in Figure.2. The resources - computational, storage and I/O - are organized as nodes of the hexagon with a local switch at the centre that interconnects these resources. Hexagons at the periphery would be primarily for I/O, whereas the ones in the core would have storage and computational resource. To further improve the connectivity, switches are directly connected to their next nearest neighbors, as shown in Figure 2, allowing any resource to reach 27 additional resources with two hops. As a last measure to further improve connectivity, every alternate switch is directly connected making each resource element reach a lot more elements with minimal number of hops.

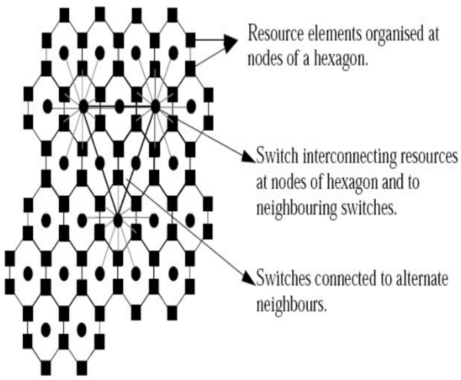


Figure.2. A honey comb structure for NOC [7]

2.2 Mesh Topology

NOC is a scalable packet switched communication platform for single chip systems. The NOC architecture consists of a mesh of switches together with some resources which are placed on slots formed by the switches.[2] Figure 3 shows NOC architecture with 16 resources. Each switch is connected to four neighboring switches and one resource. Resources are heterogeneous. A resource can be a processor core, a memory block, a FPGA, custom hardware block or any other intellectual property (IP) block, which fits into the available slot and complies with the interface with the NOC switch. We assume switches in NOC have buffers to manage data traffic.

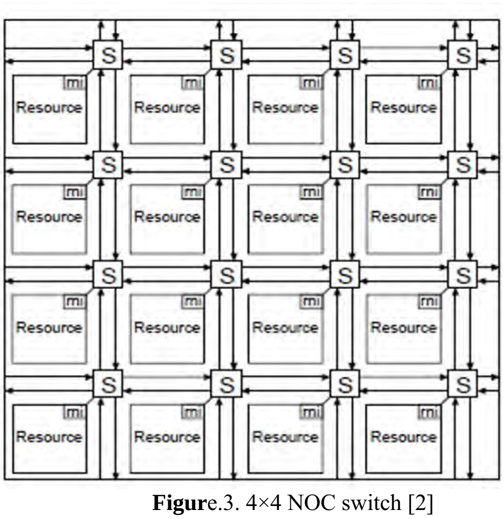


Figure.3. 4×4 NOC switch [2]

Every resource has a unique address and is connected to a switch in the network via a resource network interface (RNI). The NOC platform defines four protocol layers: the physical layer, the data link layer, the network layer, and the transport layer. The RNI implements all the four layers, whereas every switch to switch interface implements the three of four layers except physical layer. The NOC architecture also has a concept of region allows us to handle physically larger resources and can be used to provide fault tolerance. A typical NOC architecture will provide a scalable communication infrastructure for interconnecting cores. The area of multi-media is a very suitable candidate for using this high computing capacity of NOCs. NOC is a general paradigm and one needs to specialize a NOC based architecture for every application area.

III. SOPC BUILDER

SOPC Builder is a powerful system development tool. SOPC Builder enables us to define and generate a complete system-on-a-Programmable-chip (SOPC) in much less time than using traditional, manual integration methods. SOPC Builder is included as part of the Quartus II software (www.Altera.com). We used SOPC Builder to create systems based on the Nios® II processor. [3]

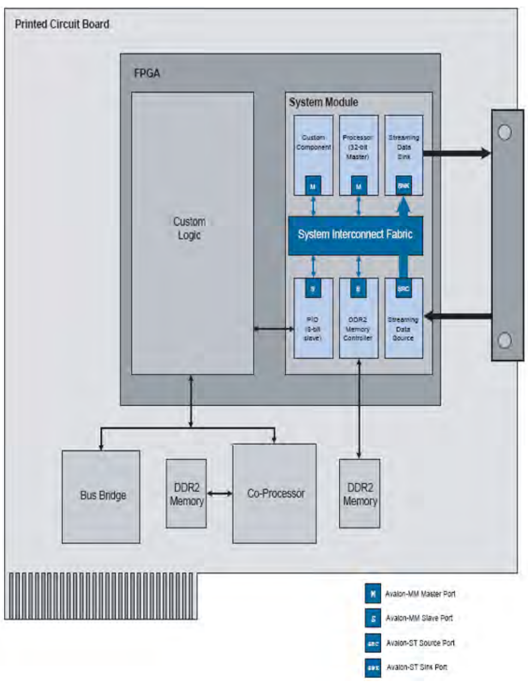


Figure.4. System example [3]

Figure.4 shows an FPGA design that includes an SOPC Builder system and custom logic modules. We can integrate custom logic inside or outside the SOPC Builder system. In this example, the custom component inside the SOPC Builder system communicates with other modules through an Avalon-MM master interface. The custom logic outside of the SOPC Builder system is connected to the SOPC Builder system through a PIO interface. The SOPC Builder system includes two SOPC Builder components with Avalon-ST source and sinks interfaces. The system interconnect fabric shown below in Figure.5 connects all of the SOPC Builder components using the Avalon-MM or Avalon-ST system interconnects as appropriate.

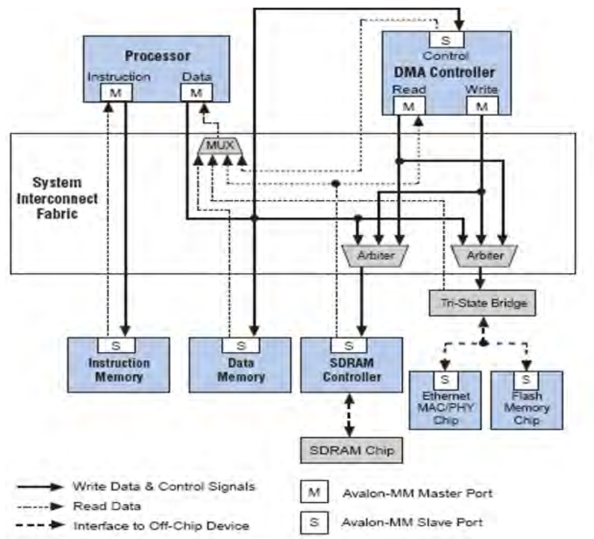


Figure.5. System interconnect fabric

The systems interconnect fabric [3] for memory-mapped interfaces are a high-bandwidth interconnects structure for connecting components that use the Avalon® Memory-Mapped (Avalon-MM) interface. The system interconnect fabric consumes minimal logic resources and provides greater flexibility than a typical shared system bus. It is a cross-connect fabric and not a tri-stated or time domain multiplexed bus. Here we describe the functions of system interconnect fabric for memory-mapped interfaces and the implementation of those functions.

3.1. Chip Planner

The Chip Planner provides a visual display of chip resources. It can show logic placement, Logic Lock and custom regions, relative resource usage, detailed routing information, fan-in and fan-out paths between registers, and delay estimates for paths.

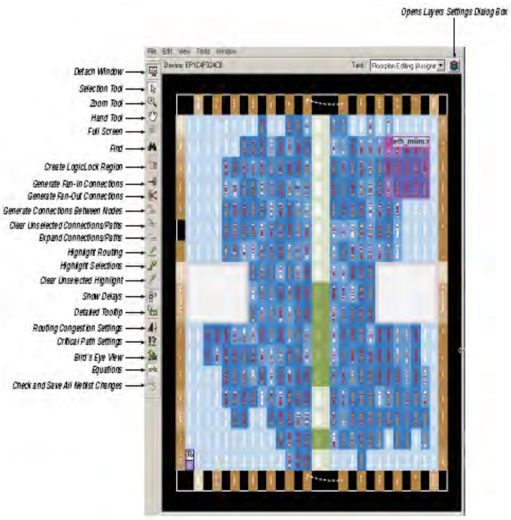


Figure.6. Chip planner tool bar from Quartus II Software

With the Chip Planner, we can view critical path information, physical timing estimates, and routing congestion. We can also perform assignment changes with the Chip Planner, such as creating and deleting resource assignments, and post-compilation changes like creating, moving, and deleting logic cells and I/O atoms. By using the Chip Planner in conjunction with the Resource Property Editor, we can change connections between resources and make post-compilation changes to the properties of logic cells, I/O elements, PLLs, and RAM and digital signal processing (DSP) blocks. With the Chip Planner, we can view and create assignments for a design floor plan, perform power and design analyses, and implement ECOs in a single tool.

3.2. Viewing Routing Congestion

The Routing Congestion view allows us to determine the percentage of routing resources used after a compilation. This feature identifies where there is a lack of routing resources. This information helps us to make decisions about design changes that might be necessary to ease the routing congestion and thus meet design requirements. The congestion is visually represented by the color and shading of logic resources. The darker shading represents greater routing resource utilization. We can set a routing congestion threshold to identify areas of high routing congestion. After selecting the Routing Utilization layer setting, click on the Routing Congestion icon on the taskbar.

3.3. Viewing I/O Banks

The Chip Planner can show all of the I/O banks of the device. To see the I/O bank map of the device, click the Layers icon located next to the Task menu. Under Background Color Map, select I/O Banks.

3.4. Generating fan-in and fan-out Connections

This feature enables us to view the immediate resource that is the fan-in or fan-out connection for the selected atom. For example, selecting a logic resource and choosing to view the immediate fan-in enables us to see the routing resource that drives the logic resource. We can generate immediate fan-in and fan-outs for all logic resources and routing resources. To remove the connections that are displayed, click the "Clear Connections" icon in the toolbar.

3.5. Highlight Routing

This feature enables us to highlight the routing resources used for a selected path or connection.

3.6. Delay Calculation

We can view the timing delays for the highlighted connections when generating connections between elements. For example, you can view the delay between two logic resources or between a logic resource and a routing resource.

3.7. Viewing Assignments in the Chip Planner

Location assignments can be viewed by selecting the appropriate layer set from the tool. To view location assignments in the Chip Planner, select the Floor plan Editing (Assignment) task or any custom task with Assignment editing mode. The Chip Planner shows location assignments graphically, by displaying assigned resources in a particular color (gray, by default). We can create or move an assignment by dragging the selected resource to a new location.

IV. RESULTS

Using SOPC Builder in Quartus II tool, we designed and simulated efficient SOPC-based Multicore System, and the results are listed in Figure 7-11.

Figure 7 shows the screenshot of the developed SOPC builder system. It is done using SOPC builder, this build system has a design of multicore system with 2 CPU's, Avalon tri state bridge, flash memory, LCD and PIO's. It's a FPGA design which includes SOPC builder system and custom logic modules. We can integrate custom logic inside or outside the SOPC builder system. In this design the custom logic modules inside the SOPC builder system communicates with other modules through an Avalon-MM-master interface. The custom logic modules outside of the

SOPC builder system is connected to SOPC system through a PIO interface.

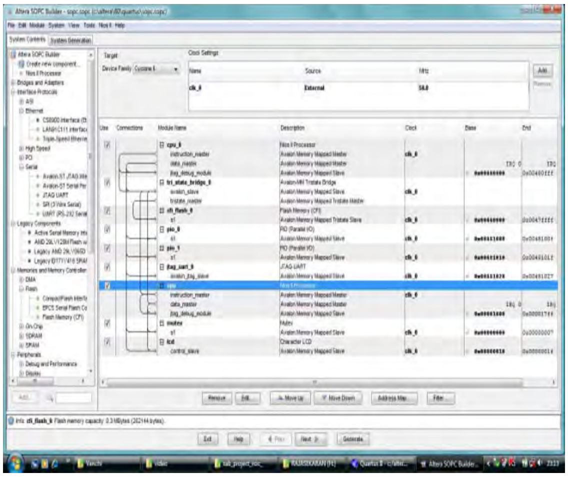


Figure.7. SOPC builder for system building

The block diagram of the symmetric file is shown in Figure 8. SOPC builder allows us to design the structure of a hardware system. The GUI allows adding components to a system configure the components and specify the connectivity. After adding and parameterize components, SOPC Builder generates the system interconnect fabric, outputs HDL files and .BDF during system generation. This .BDF file shown in Figure8 represents the top –level SOPC system for use in Quartus II.

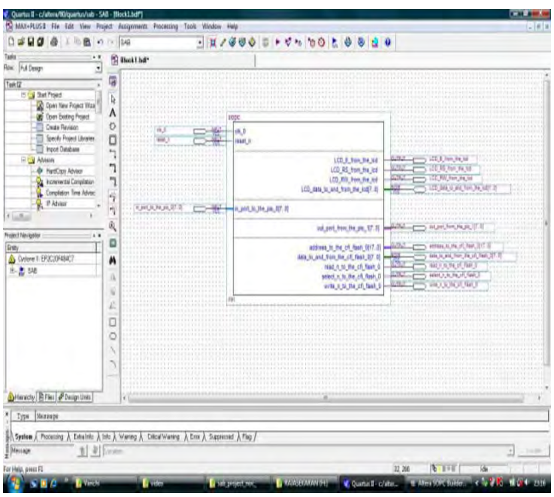


Figure.8. Block diagram of the symmetric file

The compilation result is shown in Figure 9. Once the system design is over it need to be verified whether the designed system has no errors so in order to test the system we compile our system. It shows 100% full compilation during synthesize of our SOPC system.

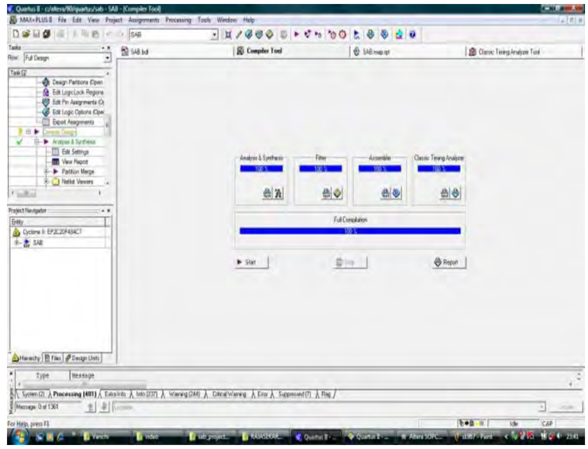


Figure.9. Compilation

The chip planner view of the compiled design is shown in Figure 10. In this screenshot we can see the build components placed in this chip planner. We can view critical path information, physical timing estimation and routing congestion. The Chip Planner uses a hierarchical zoom viewer that shows various abstraction levels of the targeted Altera device. As we increase the zoom level, the level of abstraction decreases, thus revealing more detail about our design.

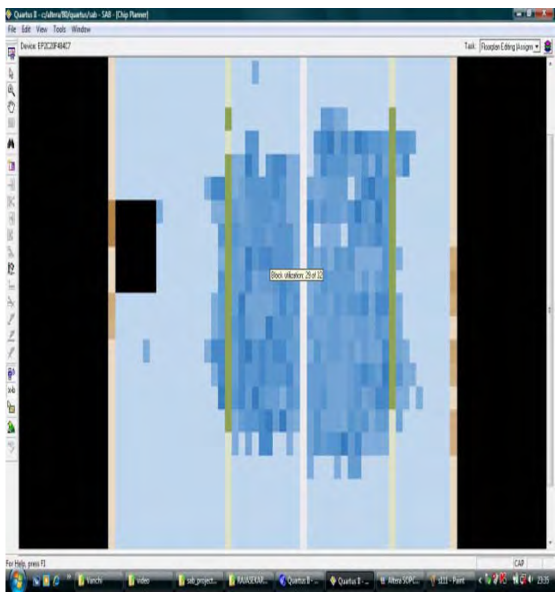


Figure.10. Chip planner view of compiled design

The routing utilization is shown in Figure 11. It allows us to determine the percentage of routing resources used after compilation. This information helps us to make decisions about design changes which is necessary to ease the routing congestion to meet the design requirements. The routing congestion is visually represented by the color and shading of logic resources. In Figure 11 we can see some areas are dark and some areas are bright. The dark regions represent greater routing resource utilization and the bright regions represent no routing congestion.

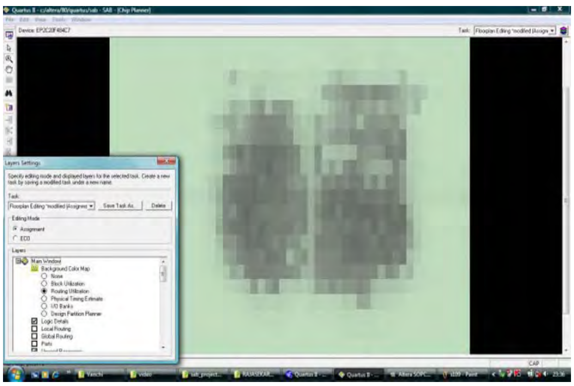


Figure.11. Routing utilization (Darker areas Show dense routing connections)

V. CONCLUSIONS AND FUTURE WORK

In this paper, a simple multicore embedded system was developed and synthesized using Altera SOPC Builder. The synthesize results demonstrate that routing will be the greater issue when it comes to the chip design. Using network-on-chip architecture a prototype system based on networking was developed to overcome the routing issue. As a result, network-on-chip micro networks are used in multicore processors to interconnect the various cores to communicate simultaneously. This leads to larger on-chip bandwidth and reduces routing congestion so that the system efficiency is enhanced. Our designed multicore system has two CPU cores which are individually optimized to the particular computational characteristics of different application fields, complementing each other to deliver high performance levels with high flexibility at reduced cost.

The research focus is shifting from implementation of NOC to investigation of its optimal use. The research problems in NOC design are identified as synthesis of communication infrastructure, choice of communication paradigm, application mapping and optimization. In the future, we will continue to design an efficient multicore SOPC with optimized timing constraints, reduced latency and improved programmability. We will also develop highly embedded, multi-core systems with more number of cores which in turn increases the system performance and many applications can run at the same time.

REFERENCES

[1] Luca Benini, Giovanni De Micheli, “Networks on Chips: A New SoC Paradigm”, *Computer*, v.35 n.1, p.70-78, January 2002.

[2] S. Kumar, A. Jantsch, J.-P. Soininen, M. Forsell, M. Millberg, J. Oberg, K. Tiensyrja, A. Hemani, “A network on chip architecture and design methodology”, *Proceedings of IEEE Computer Society Annual Symposium on VLSI*, pp. 105-112, April 2002.

[3] Quartus II Handbook, SOPC Builcer, Version 9.0, Volume 4, URL: www.altera.com/literature/hb/qts/qts_qii5v4.pdf

[4] Axel Jantesh and Hannu Tenhumen, “Networks on Chip”, *Kluwer Academic Publications*, 2003, Boston, USA.

[5] Muhammad Ali, Michael Welzl, Sybille Hellebrand, “A dynamic routing mechanism for network on chip”, *Proceedings of IEEE NORCHIP*, Oulu, Finland, 21-22, Nov. 2005.

[6] International Technology Roadmap for Semiconductors 2003, URL: http://public.itrs.net

[7] H. Tenhunen and A. Jantsch, "Networks on Chip", Springer, 1st edition, ISBN: 1402073925, Jan. 2003.

Architecture and Initial Performance Indicators of New OPCUA Automation Protocol Stack Java Implementation

Mikko Salmenperä, Mikko Salonen

Tampere University of Technology, Department of Automation Science and Engineering, P.O.BOX 692, FI 33101 Tampere, Finland Tel: 03-31154203, Fax 03-31152340, email: firstname.lastname@tut.fi

Abstract. This paper presents OPCUA binary protocol from telecommunication point of view by using OSI-network model as a basis. Presented architecture is based directly on OPCUA specifications and Java reference implementation. Practical interest is focused on OPCUA security and its feasibility for real world applications. The impact of enabling security in OPCUA is measured by benchmarking performance of current Java default implementation.

I. Introduction

Integration of industrial information systems has become a common means to increase the overall efficiency of production. Information system integration is even exceeding plant and company borders forming a collaborative network between businesses. The current trend of increasing demands posed for integration calls for unified base on which the solutions are built on. This significantly reduces the cost and time of integration in the long term.

One widely adopted solution to evolving needs of industrial integration is OPC Foundation specifications. The term OPC was originally acronym of concept OLE (Microsoft Object Linking and Embedding) for Process Control. However as OLE is no longer relevant technology for OPC UA the interpretation has been changed to Openness, Productivity and Collaboration. Over the years OPC has become the de facto standard for industrial integration and process information sharing.

The original OPC was based on Microsoft DCOM (Distributed Component Object Model) which is no longer actively developed by Microsoft. Traditional OPC is also facing integration needs that it can no longer fulfill due static information and communication models based on. These reasons have leaded the OPC Foundation to reinvent OPC replacing the core components and technologies with modern, vendor independent solutions. Backward compatibility has not been forgotten and OPC UA profiles detail use cases similar to those of legacy OPC systems. The new specification is called OPC Unified Architecture (OPC UA).

New characteristics of OPC UA are quite different from old OPC/DCOM which was meant for straight forward integration of function block based basic automation information. OPC UA focuses on providing support for complex data models and flexible address space. OPC UA abandons inherent need for Windows operating system and tight coupling with Windows security settings present in OPC/DCOM. The somewhat complex programming with COM technology is also replaced with modern software development approach. OPC UA also increases the number of possible types of integration solutions for which OPC can be used by basing on standard technologies to implement Service Oriented Architecture (SOA), namely web services (WS).

Even though OPC UA has significantly evolved from OPC/DCOM in both data model and implementation architecture backward compatibility has not been neglected. OPC Foundation has realized that clear and working way to emigrate from legacy OPC/DCOM is requirement for OPC UA success. This transformation can be accomplished by using wrapper technology for which OPC Foundation provides sample solutions.

The goal of this paper is to present structure of OPC UA communication stack as it has been implemented in JAVA reference implementation. Java implementation is closely related but not identical to that of .NET/C# implementation developed by OPC Foundation. The performance of OPC UA protocol stack is also studied based on measurements done with released version of the Java stack. Migration path to OPC UA is further studied in papers [1] and in book [3] pages 283-293. Capabilities of OPC UA in low level device integration are studied in references [4] and [5].

II. Background

This paper is based on results and expertise obtained in the OPC Unified Architecture Java reference implementation (OPCUA Java, 2006-2007) and OPC UA Information model (OPCUA Java 2, 2008 - 2009) projects. The goals of these projects were to contribute to the OPC Foundation's Early Adopter Program, develop an OPC UA Java reference implementation and assess the feasibility of the OPC UA specification in the Java programming environment.

Final results were handed over to the OPC Foundation to be published officially for OPC Foundation members.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_60,

References [1] and [2] details further results of these projects. The projects were funded by Tekes (Finnish Funding Agency for Technology and Innovation), as well as Finnish companies Metso Automation, Prosys PMS, Wapice and Kone. The Technical Research centre of Finland (VTT) and Tampere University of Technology (TUT) participated in the development of the Java implementation and Helsinki University of Technology (HUT) provided expertise.

III. OPC UA PROTOCOL

Overall functionality, architecture and complexity of OPC UA stack are rather difficult to grasp based on the specification only due to fragmentation of relevant pieces of information into 12 different specification parts. In this chapter, a brief analysis of OPC UA stack binary protocol implementation in Java is given. Architecture used in Java is closely related to C# implementation but not identical.

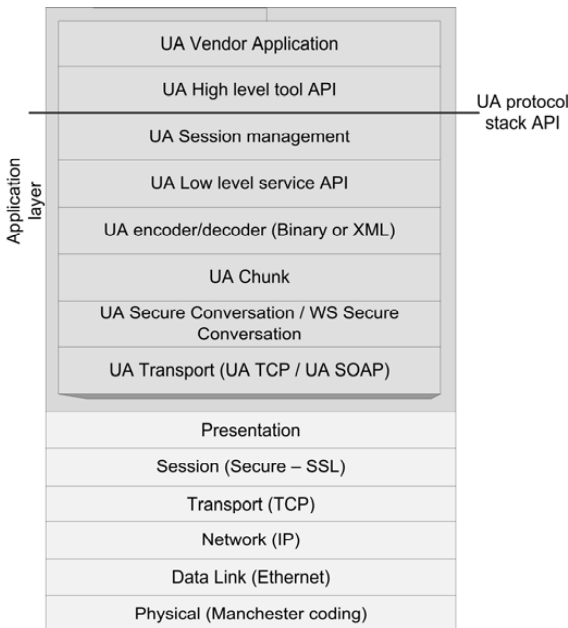


Figure 1: OPC UA Protocol stack is located on OSI models application layer

In practice, OPC UA protocol stack is located on OSI (Open Systems Interconnection Reference Model) models level 7 – Application Layer. OSI model is defined in ITU-T standard if reference [6]. UA stack contains partly similar functionality as lower levels of OSI model based network stack. This functionality could be moved away from the OPC UA stack but then the implementation would be more dependent on TCP/IP part of OSI model. OPC Foundation avoided this kind of dependency and designed the stack to be as independent and isolated from underlying transport protocols as possible. However OPC UA stack tries to utilize security and other good features from existing protocols in order to be able to utilize existing security libraries and architectures. This approach also attempts to bypass caveats present in designing secure and robust communication protocol from scratch. Presented model of OPC UA communication stack is based on the specification parts [7], [9], [10] and [11], OPC Developers Conference materials [12] and the knowledge acquired from above mentioned projects.

OPC UA stack is in itself layered as can be seen from Figure 1. Basic architecture and functionality mimics TCI/IP stack. Each layer is independent from other layers and provides certain functionality or service for the application. The lowest layer is UA Transport. It binds the stack to either TCP/IP or SOAP/Web Services protocol. OPC UA application can freely choose which transport binding it uses or even use them both. Our current Java implementation supports at the moment only binary TCP/IP binding.

UA Secure Conversation layer implements security features required by the specification. These features include various encryption and signing methods. This functionality could be moved into OSI model's session layer. This would mean basing security on simple SSL/TLS based communication used in most other current applications requiring reliable and encrypted communications. This would simplify the stack significantly and also increase security as all OPC UA protocol content would be protected by security. The current implementation leaves some header fields to plain text even if encryption is used. On the other hand by implementing security in the OPC UA stack interoperability can be promoted.

UA Chunk fragments large messages into smaller chunks that are then transferred individually over the network. On the receiver's side chunks are then assembled. Once the whole message has been reconstructed it can be passed on to upper layer in the stack. Advantage of this is that each chunk goes through the underlying security layers separately. Thus if used threaded stack each chunk can be encrypted and transferred independently and in parallel to each other. Without threading the meaning of chunks is somewhat diminished. The actual memory usage is largely unaffected by chunk mechanism as the message must be completely formed before it can be processed any further.

UA encoder/decoder layer is responsible for transforming object oriented data stored in program memory into either XML or binary presentation defined in OPC UA specification.

UA low level service API is generated service interface through which all data transfer occurs. It consists of service oriented functions defined in the specification. At the moment in Java and C# language this API is automatically generated from XML description from OPC UA web services. This API can be seen as boundary between UA stack and UA application. OPC Foundation is offering Java, C# and ANSI C stack to members. Above this API is OPC UA application either a server or a client which is self made or developed by 3rd party.

UA high level tool API signifies a need for 3rd party toolkit that expedites the development of UA applications. It provides abstraction of UA stack and possibly handles routine functionality such as certificates, sessions and access control. This layer is no required by specification but rather is implicitly present in all UA applications. Above this layer is UA Vendor application which is the actual client/server software.

IV. OPC UA BINARY PROTOCOL SECURITY FEATURES

One of the new features for the OPC UA is the security model that is a built-in property of the new protocol. The main requirement for the security model is to offer a flexible security approach as OPC UA may be used in such diverging environments. As a result, the UA applications must be able to mitigate several threat types and security requirements. OPC UA defines a secure way to transfer the data between UA applications where the generality is added with the help of UA Security Policies. The following describes the security features that are available when the UA Secure Conversation (UASC) protocol is used.

A. Secure Channel as security provider

The security of the messages that use the binary protocol can be improved by using the UASC protocol. By using the protocol, messages between the UA applications may be encrypted and signed according to the predefined encryption algorithms. As a result, the integrity and confidentiality of messages are assured.

In OPC UA, the principal factor of the security is the concept of Secure Channel, which is implemented as an independent layer to the OPC UA protocol stack. The Secure Channel is a long-term logical connection between OPC UA client application and OPC UA server that enables the confidentiality and integrity of the messages. This secure connection must be opened between the applications before the actual device and system data may be exchanged between UA applications. The Secure Channel can be opened by calling an OpenSecureChannel method specified in Secure Channel Service Set which defines the methods for the Secure Channel. The secured channel is formed according to the security handshake discussed in the next chapter. The messages that are related to the security handshake may be both encrypted and signed or they can be sent without security procedures. Naturally, the channel is secure only if the encryption has been used. Thereby, Secure Channel is created even if no security or encryption is used. [8][9]

The security of the Secure Channel is based on certificates, chosen UA Security Policy and fine-tuning of security level (Message security mode). The certificates are used to verify the parties' identities. Also, the keys contained in the certificate can be used for encryption and signature purposes. The UA Security Policy specifies the algorithms which are used to encrypt and sign the messages. These Security Policies have been defined in UA specification part seven to respond to the information security need for the different application fields, as comprehensively as possible. For example, the 128-bit encryption algorithms are sufficient to secure the normal applications at the moment, but UA also defines algorithms that use 256-bit encryption algorithms. Also, more UA Security Policies can be defined to UA according to demands. [3]

In addition to the Security Policies, the level of security in UA may be tuned using the concept of MessageSecurityMode. MessageSecurityMode defines whether the messages are signed and encrypted or only signed, before sending them to the network. There is also a 'None' mode defined, meaning that all the security features directed to messages are turned off. [3]

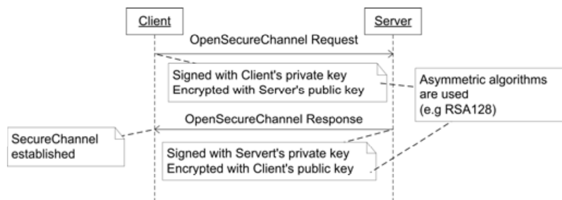


Figure 2: UA Open Secure Channel sequence

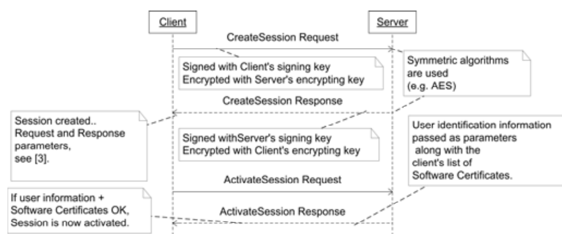


Figure 3: UA Create Session sequence

B. The Security handshake

As discussed earlier, the connections between UA applications are secured when the connection is established. This part of the UA communication is called "Security handshake" and it is performed according to the sequence which is depicted in Figures 2 and 3. The encryption and signature of messages take place by using asymmetric encryption and digital signature algorithms. The Security Handshake contains several phases: Firstly, the Secure Channel is created. Secondly, an UA Session is created and thirdly, the Session is activated. [3]

The Secure Channel may be constructed using the OpenSecureChannel service. The Session is created and activated by service methods defined in Session Service Set described in UA specification part 4 in reference [9].[9]

The OpenSecureChannel request is always made by the UA Client application that first signs the request by using its private key, after which the request is encrypted with the help of the public key to be obtained from the receiver's certificate. In this case, the receiver is the UA Server chosen to be connected to.

When an OpenSecureChannel request arrives at the UA server from the UA client (and the message has been encrypted), the server will first decrypt the message using the private key of the server. Also, the integrity of the message is verified by the server. After that, the server verifies the trust to the sender by checking the sender's certificate that is received as one part of the OpenSecureChannel request. Before the UA server may send the response message back to UA client, the server also processes other parameters of the request and calculates the derived symmetric keys to itself with the help of generated nonce. Finally, the server sends a response to the OpenSecureChannel request. The response contains along with other information, an unambiguous identifier of the channel determined by the server as well as the nonce of the server that the client program requires in order to calculate the symmetrical keys. [3][11]

The level of the information security of the response is the same as in the request. Correspondingly, the possible encryption and signature take place like in the case of the request. After the Secure Channel has been opened between the UA applications, the possible encryption and signature of messages take place with the help of symmetrical algorithms.

However, before an actual use of UA, a Session must be created and activated. This can be achieved by using the Session Service Set that defines the methods for the Session. With the help of the Session, a person may be authenticated and authorized to call the actual UA services like Browse, Read or Write. The Session creation and activation part of the security handshake has been presented in a sequence [Figure 3]. [3][9]

C. Role of Certificates

A digital certificate is a structure that associates a public key with an identity. As a result, the digital certificates may be used to identify a party that may be for example a person, an application, a product or a machine. A certificate always contains a public and private key with some other identification information depending on a certificate type. The keys combined to the certificate may be used for message encryption and signing. [3]

OPC UA uses X.509 certificates. The certificates in UA are used in three different purposes depending on the type of certificate: to verify the UA application (Application Instance Certificate), to verify UA product (Software Certificate) and to identify a user of UA client program (User Certificate). The usage of the two first mentioned certificate types is mandatory in UA, whereas the usage of the User Certificate is optional. [3]

The Application Instance Certificate of UA is a digital certificate which can be used to identify the individual UA application, thus every separate UA application (installation) requires its own application certificate. The Application Instance Certificate can also be used to sign and encrypt the messages that use the asymmetric algorithms. [3]

The Software Certificate is used to identify the particular OPC UA product. The Software Certificate contains the information about the profiles supported by the product. In practice, the OPC UA server returns in the response of the Session service to the UA client program the Software Certificates list on the basis of which the client program can conclude the supported services of the server. When a session is activated, the client program has to transmit its own Software Certificates to the UA server. The server can forbid the activating of the session, if necessary, in case the Software Certificate does not meet the demands of the server. The supplier of the UA product always delivers the Software Certificates with the product. [3][9]

The User Certificates can be used, when activating a session, to identify the user of the UA application. OPC UA also provides other opportunities to identify the user, for example the combination of username and password or web service security token instead of the User Certificate. UA also provides a possibility to use the client applications anonymously. [3] In that case, the user is not identified at all.

V. IMPACT OF ENABLING SECURITY

The usage of encryption and signature ascertain the confidentiality and integrity of the messages. However, the encryption and signature of messages require considerably more resources than those of the messages that are sent without securing them.

The Department of Automation Science and Engineering of Tampere University of Technology (TUT) began a preliminary efficiency testing concerning the security in February 2009. According to the plan, the purpose was to test the effect of different encryption algorithms on the capacity for the Java UA protocol stack. The OPC UA stack implemented using Java programming language is developed in Finland as the part of TEKES funded project, ”OPC Unified Architecture Information Modeling and Java Implementation”.

The tests are performed by sending messages between the UA applications. The messages are protected by using the several different UA Security Policies and security modes supported by the Java OPC UA Stack. The UA applications function in separate virtual machines and the measured data was collected by using the tcpdump program. In tests, the time from the client program's request (first TCP packet sent) to the time the server returns the response (first TCP packet sent) was measured. 20 requests were carried out in one measurement for each message size towards every supported security level. Every test was performed three times, resulting in total of 60 requests.

Table 1 depicts the averages calculated from the results. In figures 4, 5 and 6 the effect on the capacity of the message sizes and encryption types is seen. Figure 4 will show measurement results when a payload of 1024 bytes is transported. In Figure 5, payload was 10240 bytes, and Figure 6 outlines 102400 bytes payload transfer times.

Table 1: Averages in milliseconds

Payload size in Bytes	Rsa128 Enc & Sign	Rsa128 Sign	RSA256 Enc & Sign	None	***Effect of security***
1024	2,0	0,8	1,7	1,0	***2 X***
10240	31	1,8	27	1,7	***18 X***
102400	65	14	88	8,1	***11 X***

using the 256-bit encryption and signature algorithms. The difference with regard to the non-secure transmission of the message was about 10-fold. It is true that the message sizes that were tested are large considering the purpose of the UASC protocol, but the main target in these tests was to see the effects of the different security policies. As seen from Table 1, the differences were more distinct when the amount of data was increased.

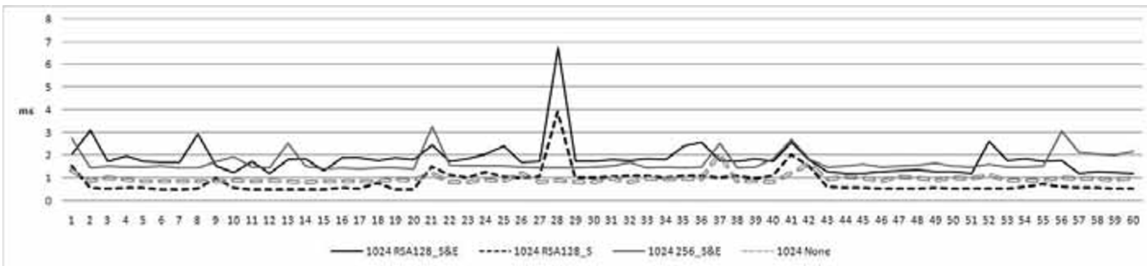


Figure 4: Packet size 1024 bytes

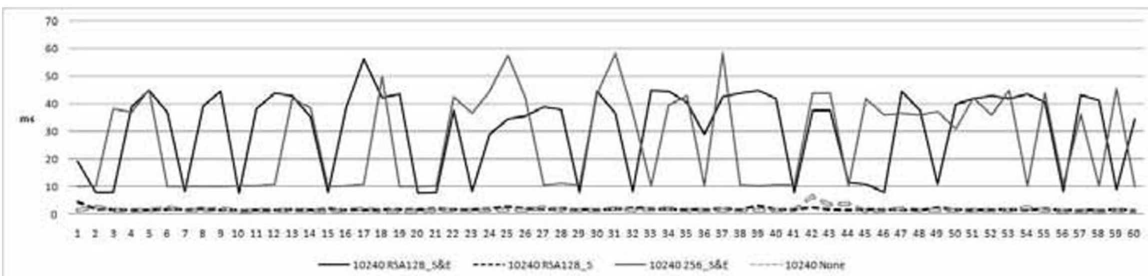


Figure 5: Packet size 10240 bytes

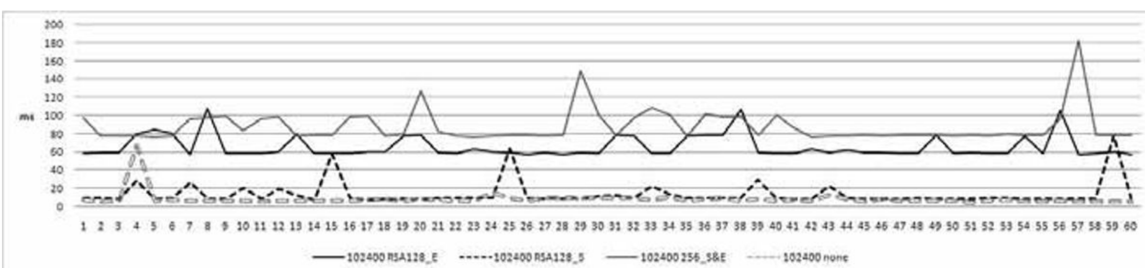


Figure 6: Packet size 102400 bytes

One can perceive from the results how much both the size of the message and the encryption algorithm used affects to the processing times. With the smallest amount of the data, the 128-bit encryption was the slowest, since the encryption and decryption of the messages took about twice a time longer than processing the message sent without any security. The biggest payload size (102400 bytes) was the slowest when

VI. Conclusions and Future Work

OPC foundation has chosen a new approach to integration. The clear and fundamental change of OPC UA compared with previous OPC specification is neutrality. OPC UA is no longer coupled to any 3rd party technology. It is also platform and programming language independent. The advantages of this choice are significant as OPC UA is targeted for integration in industrial environments, where life time of applications exceeds conventional office application by decades. On the other hand OPC specifications have grown significantly in both size and complexity.

OPC UA resides completely on OSI model's application layer even though parts of it could have been moved to lower OSI levels in OSI model. This would have reduced the size in OPC UA stack and spared resources for other important new aspects of OPC UA, such as data model. Security implementation is notoriously vulnerable to bugs and need constant updates and upkeep. It has been said that when security protocol passes initial quality assurance and audit it still needs to be used in a real world for at least five years before any creditability can be accounted to it.

Performance of OPC UA stack binary protocol implementation can be considered good and even the initial test shown in this paper clearly demonstrate that it provides a good tool for both vertical and horizontal integration. Effect of security becomes apparent with larger messages but as it can be used only where necessary there should not be major problems for even embedded environments.

REFERENCES

1. Hannelius, T., Salmenperä, M., Kuikka, S., "Roadmap to Adopting OPC UA", in 6th International IEEE Conference on Industrial Informatics (INDIN 2008), Daejeon, Korea, July 13-16 , 2008.

2. Implementing OPC UA in Java; Mikko Salmenperä, Seppo Kuikka, Jyrki Peltoniemi, Tommi Karhela, Marko Mattila, Jouni Aro, Olli Snellman, Esa Salonen; Automaatio 07 Seminar; Finnish Society of Automation; 2007; ISBN 978-952-5183-33-7

3. OPC Unified Architecture, Wolfgang Mahnke, Stefan-Helmut Leitner, Matthias Damm, 2009, Springer

4. Providing device integration with OPC UA, Hadlich, T., Industrial Informatics, 2006 IEEE International Conference on, Aug. 2006.

5. Bender K., Grossmann, D., Danzer, B, Future Device Integration, V1.1, TUM, ITM, 2007.

6. X.200: Information technology - Open Systems Interconnection - Basic Reference Model: The basic model. ITU-T Recommendation X.200. 1994.

7. OPC Foundation, 'OPC Unified Architecture Specification Part 1: Concepts', Version: Release 1.01, February 2009

8. OPC Foundation, 'OPC Unified Architecture Specification Part 2: Security Model, Version: Release 1.01, February 2009

9. OPC Foundation, 'OPC Unified Architecture Specification Part 4: Services, Version: Release 1.01, February 2009

10. OPC Foundation, '*OPC Unified Architecture Specification Part 5: Information Model'*, Version: Release 1.01, February 2009

11. OPC Foundation, 'OPC Unified Architecture Specification Part 6: Mappings, Version: Release 1.00, February 2009

12. OPC Foundation, DevCon 2007, OPC Unified Architecture, Conference Materials, OPC Foundation, 2007

Automatic Segmentation of Cardiac Cavity Images Using Collinear and Triangle Equation

Riyanto Sigit, Mohd. Marzuki Mustafa, Aini Hussain, Oteh Maskon, and Ika Faizura Mohd Noh
Dept. of Electrical, Electronic and Systems Engineering,
Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia
Bangi, Malaysia
riyanto@eepis-its.edu, marzuki@eng.ukm.my, aini@eng.ukm.my

Abstract- **Automatic segmentation of cardiac cavity images using collinear and triangle equation algorithms to detect and reconstruct the boundary of cardiac cavity is afforded in this paper. Firstly, high boost filtering is used to enhance the high frequency component without affecting the low frequency component. Next, to eliminate noise and convert the image into binary format, the morphological and thresholding operators are applied to the image. Then, using negative Laplacian filter followed by region filtering, the edge detection is performed. Finally, the collinear and triangle equations are used to detect and reconstruct the more precise cavity boundary. Results obtained showed an improvement in the detection boundary of cardiac cavity which implies the suitability of the proposed technique in performing the boundary detection task of cardiac cavity from echocardiographic images.**

Keywords: **Cardiac cavity, high boost filter, morphology, negative Laplacian, region filter, collinear and triangle equation.**

I. Introduction

The two-dimensional (2D) short axis echocardiography images have been used by medical practitioners to diagnose patient cardiac condition by studying the cardiac cavity. As such, various researches and methods have been conducted to develop computerized technique to detect cardiac cavity [1-9]. Although many techniques have been developed, there is still room for innovation and development of such methods and algorithms. For cardiac activity detection, researchers have used either the short axis images [1-6] or long axis images [7&8] implemented in semi-automatic mode [1, 4 & 6] or fully automatic [3 & 7]. In addition, active contour models or snakes have also been used to segment cardiac cavity [5 & 8]. J. W. Klinger et al in [1] have applied segmentation of echocardiography images using mathematical morphology.

On the other hand, Laine and Zong in [2] have used border identification that depended on the shape modeling and border reconstruction from a set of images. Detection of left ventricular endocardium using ternary threshold method was introduced by Ohyama et al [3], in [1 & 4] the semi automatic detection method was employed and the multiple active contour models for cardiac boundary detection by Chalana et al is presented in [5]. In a more recent work, watershed pre-segmentation snake was used for boundary detection and to find its center point [6 & 7]. Of late, applied radial-search was used for segmentation of cardiac cavity [8]. Accordingly, we have devised a technique which involves the use of collinear and triangular equations. Our proposed technique employs the high boost filter to enhance the high frequency component while keeping the low frequency component unaffected. This is followed by applying the morphological and thresholding operators to the image to eliminate noise and convert it to the binary format. Next, Laplacian filter and region filtering are applied in order to perform the edge detection process.

The final step is the implementation of the collinear and triangle equation which is used to detect and reconstruct an improved cardiac activity border. It is therefore, the aim of this paper to present our proposed technique which we believe can be used to perform segmentation of cardiac cavity very precisely. The following section describes the development of the proposed technique further and outcomes of the implementation are shown in the results and discussion section.

II. Proposed Automatic Segmentation Method

The schematic diagram of the proposed technique is shown in Fig.1. As shown, the collinear and triangle equations implementation represents the last step in our proposed automatic segmentation technique to detect edge and reconstruct the border. Prior to it, the echocardiography image is subjected to a 3-step procedures involving high boost

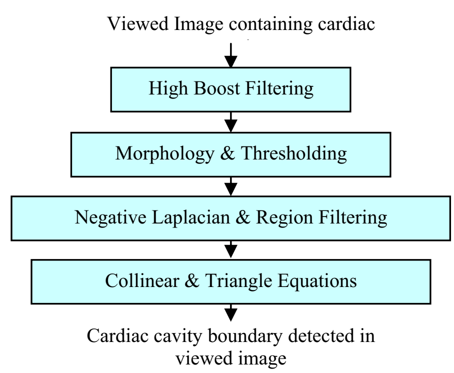


Fig.1. Schematic Diagram of the algorithm

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_61, © Springer Science+Business Media B.V. 2010

filtering, morphological and thresholding and a double filtering using the Laplacian and region filters.

A. Cardiac Short Axis Image View

There are two main standard views used in cardiac imaging that is the short axis and long axis. In this research, we have used the short axis images to test out our proposed segmentation technique. An example of the image comprising the left ventricular cardiac cavity from echocardiography is shown Fig. 2.

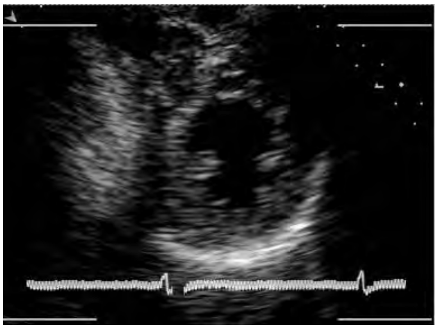


Fig.2. Screen-shot of the short axis left ventricular cardiac cavity image from echocardiography

B. High Boost Filter

The first step in our proposed technique is the implementation of the high boost filter algorithm [6], which uses the spatial mask shown in Fig. 3. By using the spatial mask, the high-boost filter is used to enhance high frequency component without affecting the low frequency components.

-0.1111	-0.1111	-0.1111
-0.1111	9.89	-0.1111
-0.1111	-0.1111	-0.1111

Fig.3. Mask used for the high boost filter.

C. Morphology and Thresholding

After image enhancement using high boost filter, the enhanced image is subjected to the morphological operation and thresholding. Morphological operation involves the use of the opening and closing algorithms [1]. The main function of the algorithms is to reduce speckle noise in the cardiac cavity image. The opening algorithm involves eroding image A by B and dilation by B. The mathematical notation of the opening algorithm is shown in equation 1.

$$A \circ B = (A \ominus B) \oplus B \tag{1}$$

where $\ominus$ and $\oplus$ denote erosion and dilation, respectively.

On the contrary, the closing algorithm is when image A is dilated and eroded by B. The mathematical notation of the closing algorithm is shown in equation 2.

$$A \bullet B = (A \oplus B) \ominus B \tag{2}$$

D. Negative Laplacian & Region Filter

Next, the negative Laplacian filter is implemented. It is a derivative filter that is used to find areas of rapid change (i.e. edges) in an image. There are different ways to find an approximate discrete convolution kernel that can approximate the effect of the Laplacian. A possible kernel is as shown in Fig.4 below.

0	1	0
1	-4	1
0	1	0

Fig. 4. Kernel used for negative Laplacian.

Following negative Laplacian filtering, the image is then subjected to region filtering. The main function of region filter is to eliminate small contours. The region filter scans the contour and calculates the area of each contour. Regions with area that is smaller than the pre-determined threshold are eliminated from the contour [6]. The threshold value is set to 25 pixels, which was empirically determined.

E. Collinear & Triangle Equation

Next, the collinear equation algorithm is used to find centroids of all contours by keeping and deleting some contours so that the resulted contour is closer to the actual boundary. All existing contour centroids are computed using equation 3 shown below.

$$Centroid(C) = \left\{ \frac{\sum_{k=1}^{n} X_k}{n}, \frac{\sum_{k=1}^{n} Y_k}{n} \right\} \tag{3}$$

The collinear equation is applied from center of the boundary to contour centroids by finding the slope and intercept using equation 4, 5 and 6.

$$y = wx + b \tag{4}$$

$$Slope\ (w) = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \tag{5}$$

$$Intercept\ (b) = \bar{y} - w\bar{x} \tag{6}$$

If the slope of the contour intersects with the inner boundary, the contour is maintaned and removed if it intersects with the outer boundary. Fig. 5 shows a triangle, where *A, B, C* represents the corners, and *a, b, c* are the distances between corners.

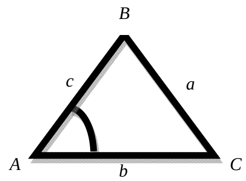


Fig.5. A triangle

If *a, b,* and *c* are known, then the angle of each corner can be calculated using equation 7 and 8.

$$a^2 = b^2 + c^2 - 2bc(\cos A) \quad (7)$$

$$A = a\cos((b^2 + c^2 - a^2)/2bc) \quad (8)$$

After thresholding operation, the resultant cardiac cavity image will consist of a closed border and some open borders. To close the borders one can utilize the small corner angle using the triangle equation. Accordingly, we utilized the OpenCV Library, which is a library for computer vision to get the boundary contour and reconstruct it. To do so, some modification is required and as such we have made modification to this library. In the OpenCV library, a contour that can be stored inside the memory storage is known as a sequence. A sequence in OpenCV is actually a linked list [7].

The `cvFindContours` function is used to retrieve contours from the binary image and to return the number of retrieved contours. The function `cvSeqPop is used to` remove an element from the sequence and `cvSeqPopFront` function, on the other hand, is used to remove an element from the beginning of the sequence. As such, one can perform contour cutting by using the `cvSeqPop` function and `cvSeqPopFront` function.

However, to do so it is crucial that the initial center boundary of the cardiac cavity is determined by using the `cvFindContours` function. Once the center is determined, the next step is to cut the contour radius that is greater than 0.5 of the image width size and then connect two end points to each other. This process is repeated until all small contours are eliminated.

III. Results And Discussion

The result of the high boost filter implementation is shown in Fig. 6 followed by the results after morphological operation with thresholding in Fig. 7. In can be seen that high boost filter implementation yields an enhanced image of the cardiac cavity which is further improved by using morphological operations with thresholding. Result from the negative Laplacian filter implementation is shown in Fig. 8 while result after region filtering is shown in Fig. 9. Accordingly, Fig. 10 depicts the result from collinear equation computation where a contour is maintained if the slope of the contour intersects with the inner boundary and removed if it intersects with the outer boundary.

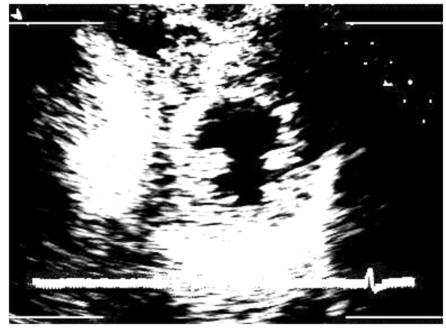


Fig. 6. Sample result from high boost filter implementation

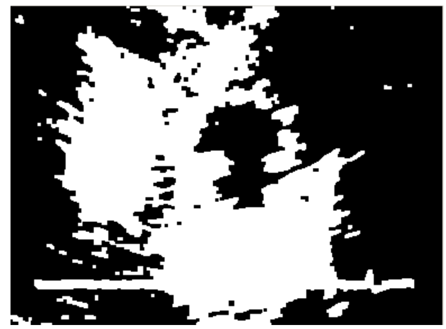


Fig. 7. Sample result after morphological operations with thresholding

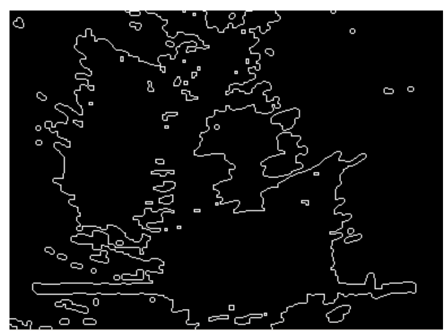


Fig.8. Sample result from negative Laplacian filter implementation

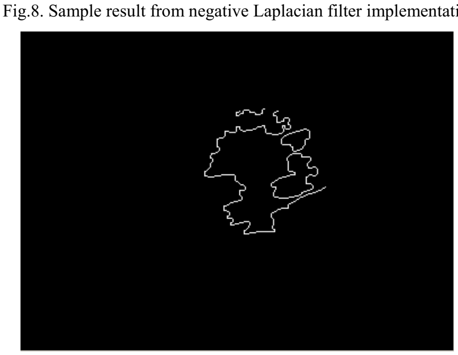


Fig.9. Sample result from region filter implementation

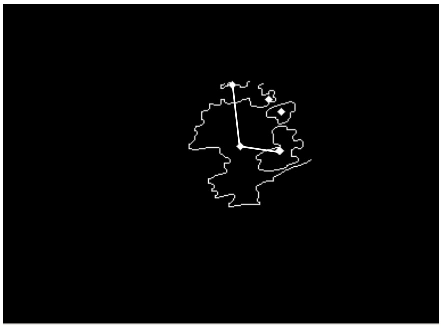


Fig.10. Sample result from collinear implementation

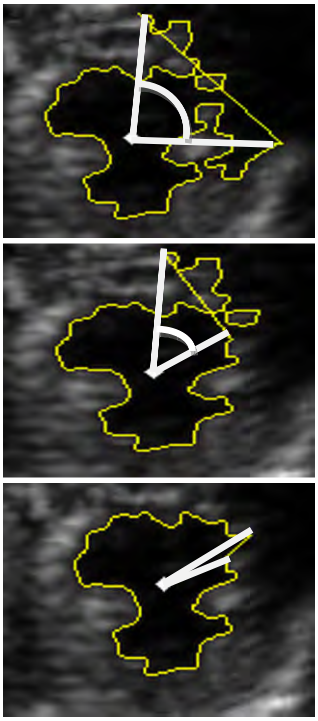


Fig.11.The step-by-step results to eliminate small contours using the OpenCV library functions.

Step-by step results using the triangle equation to eliminate small contours by small corner angle computation are shown in Fig. 11(a), (b) and (c). Figure 11(c) depicts the final cardiac cavity boundary that has been successfully detected and precisely reconstructed.

All these steps are then validated by testing with images from a video recording that consists of 9 frames. The typical image size used is 320 pixels wide and 240 pixels high. Fig. 12 shows the reconstructed cardiac cavity boundaries of all nine frames. It can be seen our proposed method successfully detects and trace the boundaries of all cardiac cavity in the video as it changed from large to small.

IV. CONCLUSION

A practical solution to perform automatic segmentation of cardiac cavity using collinear and triangle equation has been described. The proposed method successfully detects and reconstructs precise boundary of cardiac cavity which can help in the cardiac investigations by medical practitioners.

ACKNOWLEDGMENT

The authors would like to thank the government of Malaysia for the funding of this research through research grant contract number UKM-GUP-TKP-08-24-080.

REFERENCES

[1] J. W. Klinger, C. L. Vaughan, and T. D. Fraker, "Segmentation of echocardiographic images using mathematical morphology," IEEE Trans.Biomed. Eng., vol. 35, 1988, pp. 925–934.

[2] A. Laine, and X. Zong, "Border Identification of Echocardiograms via multiscale edge detection and shape modeling," Proc. of the IEEE Int. Conf. on Image Processing, vol. 3, Sep. 1996, pp. 287 - 290 vol.3.

[3] W. Ohyama, T. Wakabayashi, F. Kimura, S. Tsuruoka, and K. Sekioka, "Automatic Left Ventricular Endocardium Detection in Echocardiograms Based on Ternary Thresholding Method", in 15th International Conference on Pattern Recognition (ICPR'00), Barcelona, Spain, 2000, pp. 320-323.

[4] Maria do Carmo dos Reis, Adson F. da Rocha, Daniel F. Vasconcelos, etc, "Semi-Automatic Detection of the Left Ventricular Border ", 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 2008.

[5] V. Chalana, D. T. Linker, D. R. Haynor and Y. Kim, "A Multiple Active Contour Model for Cardiac Boundary Detection on Echocardiography Sequences", IEEE Trans. on Medical Imaging, vol.15, no. 3, June 1996.

[6] Cheng, J, Foo, S.W. & Krishnan, S. M. "Automatic Detection of Region of Interest and Center Point of Left Ventricle using Watershed Segmentation," IEEE Int. Symposium on Circuits and Systems, vol. 1, n. 2, May 2005, pp. 149-151.

[7] Cheng, J, Foo, S.W. & Krishnan, S. M. "Watershed-Pre-segmented Snake for Boundary Detection and Tracking of Left Ventricle in Echocardiographic Images," IEEE Trans. on Information Technology in Biomedicine, vol. 10, n. 2, 2006, pp. 414-416.

[8] Lacerda SG, da Rocha AF, Vasconcelos DF, de Carvalho JL, Sene IG Jr & Camapum JF, "Left Ventricle Segmentation in Echocardiography Using a Radial Search Based Image Processing Algorithm", 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 2008

[9] M. Kass, A.Witkin, and D. Terzopoulos, "Snakes: Active contour models", presented at the Int. Conf. Computer Vision, ICCV'87, London, U.K., 1987.

[10] Gary Bradski and Adrian Kaehler, "Learning OpenCV", O'Reilly Media, September 2008.

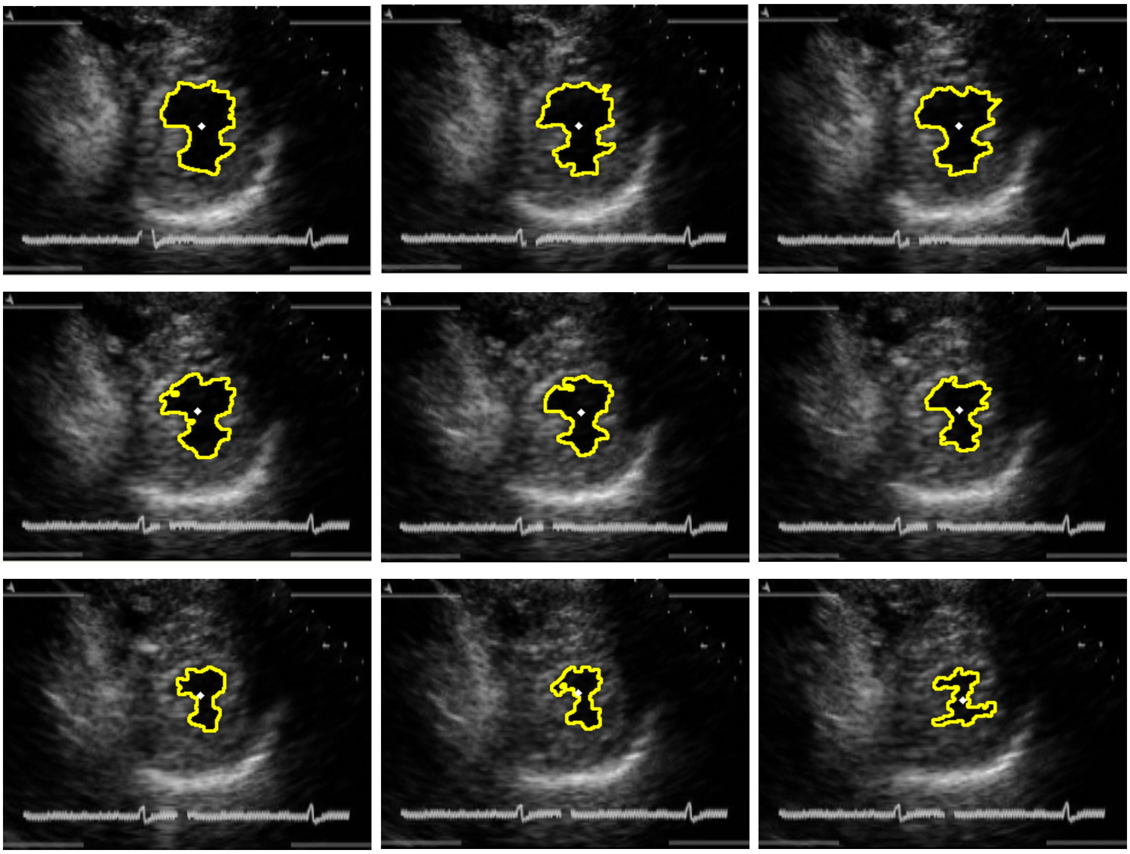


Fig.12. Frame-by-frame border detection of cardiac cavity in echocardiograph image

Both Side More-Output Transmission with Harmonic Gear

Ing. Jozef Haľko, PhD.
Faculty of manufacturing Technologies,
TU Košice with seat in Prešov, Štúrova 31,
08001 Prešov, Slovak republic
jozef.halko@tuke.sk

***Abstract* - The paper presents the design of the principle of multiple output both side differential transmissions with integrated harmonic transmission. The presented principle enables more coincidental coaxial outputs with the alternative of input and output interchange.**

INTRODUCTION

Department of technological equipment design of the FVT of the TU in Košice within the process of searching for alternatives to the so-called high precision transmissions of the constructions based on cycloidal teeth with the relatively complicated transmission of movement from satellite to centric has aimed it research at the new principles for design with relatively simple transmission of movement as well as presented transmissions regarding the number and range of transmission ratios.

The principle of integrated harmonic transmission with the multiple both side outputs has been suggested.

The paper also presents the kinematic analysis of the designed construction with transmission ratios for the certain levels.

I. PRINCIPLE OF MULTIPLE BOTH SIDE OUTPUT TRANSMISSION WITH INTEGRATED HARMONIC TRANSMISSION

Modern standard harmonic transmissions are usually designed as „one-way“ with the only input and output without the possibility of their interchange.

The suggested multiple output transmission with integrated harmonic transmission is considered to be the new design with the variants of one input and three outputs with one braking of one of the transmission members. With the two levels of latitute, the transmission can operate as a differential. Output shafts are possible on both side.

From the design principle it follows that alternative transmission from input shaft to the three output shafts (members) is achieved by means of one one-level transmission integrally interconnected with the two two-level transmissions within one construction of the whole transmission mechanism (fig 1). Transmission of revolutions and performance for the suggested design is possible in two alternatives with the possible determination of corresponding transmission ratios.

I.I THE FIRST ALTERNATIVE

Transmission of revolutions and performance is possible through the input eccentric shaft I, where fulcrum wheel „2“ and „7” is located with gearing z_2 and z_9 and firmly connected gearing z_5 and z_7, meshing with gearing z_6 eventually z_8 connected with the output shafts II, V.

The next output is possible either through the flexible member (disk wheel) „3“ to the output shaft III, or through the crown wheel „4“ connected with the output shaft IV. The second output is done through the flexible member „3“ with the internal gear z_{3a} and external gear z_{3b}.

Thus, gearing z_2 and z_9 of the wheel „2“ and „7” meshes with internal gear z_{3a} of the flexible member „3“ and consequently the flexible member by means of external gear z_{3b} meshes with the internal gear z_4 of the non-flexible member of the crown wheel„4“. If the crown wheel „4“ is braked, corresponding performance of the third output is transmitted to the shaft III by means of the flexible member „3“. If the flexible member „3“ is braked , performance is transmitted by means of the member „4“ to the shaft IV.

I.II THE SECOND ALTERNATIVE

To transmit revolutions and performance, the driveshaft II can be used. Then shafts I, III, V can be the input members, if the member „4“ eventually I, IV, V is braked, if the member „3“ is braked. Analogically it is possible to determine output shafts, if shaft V is a driveshaft.

Figure 2 shows kinematic ratios between memebers „2“ a „3“ and consequently transmission ratios for the first and the second alternatives have been derived.

II. TRANSMISSION RATIOS FOR THE FIRST ALTERNATIVE – SHAFT I IS A DRIVESHAFT

II.I TRANSMISSION RATIOS WHEN MEMBER „3“ IS BRAKED, $n_3 = 0$

<u>a) Transmission ratio between shaft I and gear wheel „2“ if the flexible member „3“ is braked</u>

The analogue relations are also between shaft I and gear wheel „7”.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_62, © Springer Science+Business Media B.V. 2010

From the kinematic ratios shown in fig.2 it follows that for the velocity vector at point „A" the following is valid:

$$\bar{v}_{A2O} = \bar{v}_{A3O} \quad (1)$$

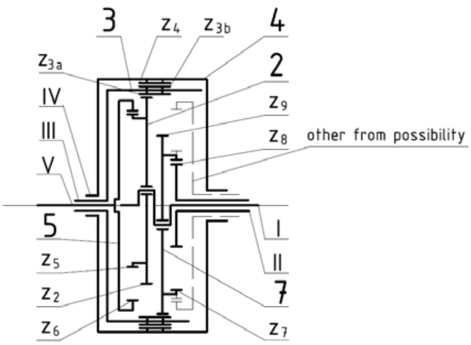


Fig. 1 Principle integrated harmonic gears of mechanism

(Two step gear – authorship certificate Ing. Jozef Haľko, PhD., number 3937, Slovakia)

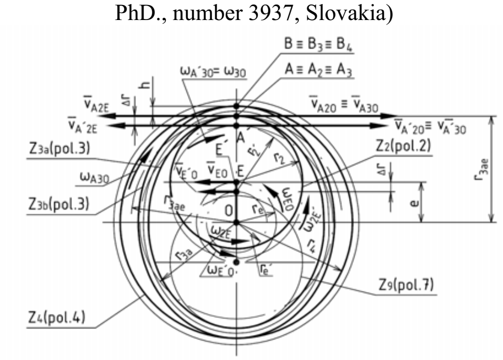


Fig. 2 Kinematic scheme

At point „A" the pitch circle of the gearing z_2 of the member „2" touches the pitch ellipse of gearing z_{3a} of the deformed member „3", basic shape of which is circle with pitch circle radius r_{3a}, of the gearing z_{3a}. At the same time, meshing of gearing z_4 of the member „4" with gearing z_{3b} of the member „3" at point „B" occurs. At this point the pitch circle touches gearing z_4 with the pitch ellipse of the gearing z_{3b} of the deformed member „3". It follows that big (main) axle of pitch ellipse of gearing z_{3a} at the point „A" has the size r_{3ae}. It follows from the above as well as from fig. 2 that:

$$r_{3ae} = r_e + r_2 \qquad r_{3ae} = r_{3af}$$

where:

r_e - axle eccentricity of the member „2" rotation

r_2 - radius of pitch circle of gearing z_2 of the member „2".

Radius r_{3ae} is smaller than r_4 by a value h (see fig. 1)

Thus: $h = r_4 - r_{3ae}$

Value h usually follows from the body „3" construction based on pre-determined requirements for a transmission. From the vector kinematics for velocity at the point „A 2" regarding point „O" the following is valid:

$$\bar{v}_{A2O} = \bar{v}_{A2E} + \bar{v}_{EO} = \bar{v}_{A3O} \quad (2)$$

From the equation (1) it follows that at the point A the following is valid:

$$\omega_{A2O} = \omega_{A3O} \quad (3)$$

After substitution the velocity vectors at the point A

$$\bar{v}_{A3O} = \omega_{A3O} . r_{3ae}, \bar{v}_{A2E} = \omega_{2E} . r_2, \ \bar{v}_{EO} = \omega_{EO} . r_e$$

and if we take into consideration the supposed direction of revolution regarding fig. 2 the following is valid:

$$\omega_{A3O} . r_{3ae} = \omega_{2E} . r_2 + \omega_{EO} . r_e \quad (4)$$

Regarding the fact that body „3" is in the part of gearing z_{3a}, z_{3b} flexible, ω_{A3O} it does not always equal the angle velocity of the body „3", thus $\omega_{A3O} \neq \omega_{3O}$. Angle velocity of the non-deformed body „3" equals angle velocity of the input shaft III. With constant $\omega_I = \omega_{EO}$ also ω_{3O} = constant. When deriving the transmission ratio ω_{3O} between shaft I and wheel „2" it is necessary to follow from circumferential velocity $v_{A'3O}$ at point A' on the pitch circle of gearing z_{3a} of the non-deformed member 3. Thus $v_{A'3O} = \omega_{A'3O} r_{3a}$, with $\omega_{A'3O} = \omega_{3O}$.

If we consider that body „3" (flexible member) is braked in the body of transmission then when deriving the transmission ratio between input shaft and wheel „2" we follow form the fact that gear wheel „2" when shaft I with the certain slew angle performs the same mesh trajectory of deformed gear member „3" by gearing z_{3a} as in case of non-deformed member „3" with the other corresponding eccentricity $r_{e'}$, that is smaller than eccentricity r_e by a value of deformation Δr (see fig. 2). Based on the above mentioned, it is possible to write the following for the velocity vector at the point A', which represents the meshing the gear wheel „2" with non-deformed member „3" with the gearing z_{3a}:

$$\bar{v}_{A'2O} = \bar{v}_{A'2E} + \bar{v}_{E''o} = \bar{v}_{A'3O} \quad (5)$$

after substitution:

$$\omega_{A'3O} . r_{3a} = \omega_{2E} . r_2 + \omega_{E'O} . r_{e'}$$

and the following is valid:

$$r_{e'} = r_{3a} - r_2 \ , \ \omega_{2E} = \omega_2, \ \omega_{E'O} = \omega_{EO} = \omega_1$$

after substitution the following is valid:

$$-\frac{\omega_I}{\omega_2} = \frac{r_2}{r_{3a} - r_2} \quad (6)$$

Expression $\frac{\omega_I}{\omega_2}$ is the searched transmission ratio. Let us denote it by u^3_{I-2}, then we can write:

$$u^3_{I-2} = -\frac{r_2}{r_{3a} - r_2} \quad \text{or} \quad u^3_{I-2} = \frac{r_2}{r_2 - r_{3a}} \quad (7)$$

where u^{3}_{I-2} - is transmission ratio between shaft I and wheel „2“ when member „3“ is braked.

If we substitute $r_{3a} = \frac{m z_{3a}}{2}$ and $r_2 = \frac{m z_2}{2}$, (m- module gearing z_2,z_{3a}) then transmission ratio u^{3}_{I-2} is as follows:

$$u^{3}_{I-2} = \frac{z_2}{z_2 - z_{3a}} = \frac{\omega_I}{\omega_2} \qquad (8)$$

<u>b) Transmission ratio between the shafts I and IV when member „3“ is braked.</u>

Output shaft V has internal gear wheel.
From the equation (6) and (8) for kinematic ratios of gear wheels z_2 and z_3 of the harmonic transmission for angle velocity of the internal gear wheel „2“ it follows

$$\omega_2 = \omega_I \frac{r_1 - r_2}{r_2} \qquad (9)$$

If annual gears z_2 and z_8 are on the same body of the gear wheel the following is valid: $\omega_2 = \omega_8$.
Eccentricity e at the same time equals the radius of revolution of the wheel axis „2“ (of the point E) around the point „O“ of the central axis of the transmission.
Thus: $e = r_e = r_6 - r_5 = r_{3a} - r_2$. (10)

From fig. 3 it follows : $\bar{v}_{C6O} = \bar{v}_{6E} + \bar{v}_{EO}$, (11)
At the same time the following is valid: $v_{C6O} = v_{C5O}$,
then $v_{C6O} = r_6\omega_6$, $v_{C6E} = r_3\omega_5$, $v_{EO} = r_e\omega_{EO}$, and $\omega_{EO} = \omega_I$.
After substitution and regarding the selected revolution according to fig. 3 we will obtain:

$$r_6\omega_6 = r_5\omega_5 + r_e\omega_I \qquad (12)$$

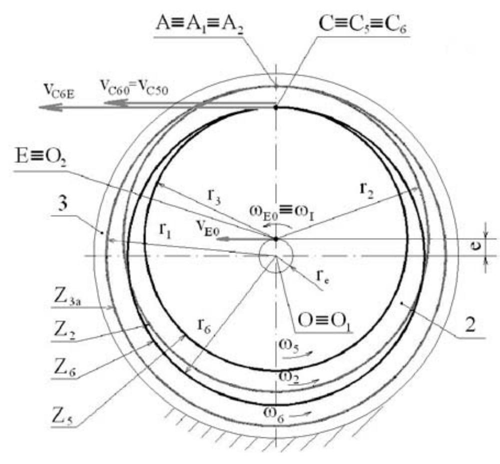


Fig. 3 Kinematic scheme of the second level transmission. Output shaft with the external gear wheel.

If $\omega_5 = \omega_2$ (gearings z_5 and z_2 are firmly gripped on the wheel „2“) then we can substitute from the equation (9) $\omega_2 = \omega_5$ to the equation (12). From the transmission design it follows that angle velocity of the shaft V equals the angle velocity of the gear wheel z_4, because of the firm connection with the shaft „V“. Thus $\omega_6 = \omega_V$.

After substitution we will obtain:

$$r_6\omega_V = r_5\left(\omega_i \frac{r_2 - r_{3a}}{r_2}\right) + (r_6 - r_5)\omega_I$$

Ratio $\frac{\omega_I}{\omega_V}$ is the transmission ratio between shaft I and V when member „3“ is braked. Let us denote it by u^3_{I-V} and angle velocity ω_V by ω^3_V (index 3 – braked member „3“ , and consequently we will obtain the following:

$$u^3_{I-V} = \frac{\omega_I}{\omega^3_V} = \frac{r_6 r_2}{r_5(r_2 - r_{3a}) + r_2(r_6 - r_5)} \qquad (13)$$

After the radius through teeth number has been defined:

$$u^3_{I-V} = \frac{\omega_I}{\omega^3_V} = \frac{z_6 z_2}{z_5(z_2 - z_{3a}) + z_2(z_6 - z_5)} \qquad (14)$$

According to the equation (14) is to achieve large size gear ratio u^{3}_{I-IV} . Normally is possible to achieve then divider the equation (14) equal zero. Then $u^{3}_{I-IV} = \infty$ (input shaft “I” turn, output shaft “V” stay).
Transmission ratios between the shaft I and II with the corresponding numbers of teeth for gearing are of the analogical form.

<u>c) Transmission ratio between the shaft I and IV when member „3“ is braked.</u>

For this transmission, analogical relationship known form harmonic transmissions can be derived. Thus, the following is valid:

$$u^{3}_{I-IV} = \frac{z_4}{z_4 - z_{3b}} \qquad (15)$$

II.II TRANSMISSION RATIOS WHEN MEMBER „4“ IS BRAKED AND $n_4 = 0$.

<u>a) Transmission ratio between the shaft I and III</u>
For the transmission ratio the following relationship can be derived:

$$u^{4}_{I-III} = \frac{\omega_I}{\omega^4_{III}} = \frac{z_{3b}}{z_{3b} - z_4} \qquad (16)$$

where ω_{III}^4 - is the angle velocity of the shaft III when member„4“ is braked.

b) Transmission ratio between the shaft I and V

For the angle velocity of the shaft V the following is valid: $\omega_V^4 = \omega_V^3 + \omega_{III}^4$ (17)

where

ω_V^4 - is the angle velocity of the shaft V when member „4“ is braked

ω_V^3 - is the angle velocity of the shaft V when member „3“ is braked

Let us denote transmission ratio between the shaft I and V when member „4„ is braked by u_{I-V}^4, then

$$u_{I-V}^4 = \frac{\omega_I}{\omega_V^4} = \frac{\omega_I}{\omega_V^3 + \omega_{III}^4} \quad (18)$$

After expression ω_V^3 from the equation (14) and ω_{III}^4 from the equation (16) and after substitution to the equation (18) we will obtain:

$$u_{I-V}^4 = \frac{z_6 z_2 z_{3b}}{\left[z_5\left(z_2 - z_{3a}\right) + z_2\left(z_6 - z_5\right)\right] z_{3b} + \left(z_{3b} - z_4\right) z_6 z_2}$$

Relationship foR transmission ratio u_{I-II}^4 between the shafts I and II when member 4 is braked is of the analogical form , although when the corresponding number of teeth of the gearing are substituted.

III. TRANSMISSION RATIOS FOR THE SECOND ALTERNATIVE – SHAFT II IS THE DRIVESHAFT

III.I. IF THE WHEEL „4“ IS BRAKED AND $n_4 = 0$.

a) Transmission ratio between the shafts II and I

$$u_{II-I}^4 = \frac{1}{u_{I-II}^4}$$

b) Transmission ratio between the shafts II and III

$$u_{II-III}^4 = \frac{u_{I-III}^4}{u_{I-II}^4}$$

c) Transmission ratio between the shafts II and V

$$u_{II-V}^4 = \frac{u_{I-V}^4}{u_{I-II}^4}$$

III.II. TRANSMISSION RATIOS WHEN MEMBER „3“ IS BRAKED AND $n_3 = 0$

a) Transmission ratio between the shafts II and I

$$u_{II-I}^3 = \frac{1}{u_{I-II}^3}$$

b) Transmission ratio between the shafts II and IV

$$u_{II-IV}^3 = \frac{u_{I-IV}^3}{u_{I-II}^3}$$

c) Transmission ratio between the shafts II and V

$$u_{II-V}^3 = \frac{u_{I-V}^3}{u_{I-II}^3}$$

CONCLUSION

The objective of the paper is to present principally new design of multiple output both side transmission with the possibility of input and output interchange with the alternatives of the possible three outputs from the given transmission mechanism. Based on the kinematic scheme, corresponding transmission ratios have been derived. Within the further research, force ratios analysis and possible loadings of the mentioned mechanism, eventually the other corresponding solutions [1], [4], [9] will be given great attention. This contribution forms a part of the solution of the grant task VEGA 1/4156/07 and KEGA 3/6279/08.

REFERENCES

1. HAĽKO, J. – KLIMO, V.: Dvojstupňový viacvýstupový prevod. In: *Sborník 45. mezinárodní konference kateder částí a mechanismů strojů: 7. – 9. září 2004, Blansko,* Česká republika, Brno: VUT, 2004, 6 s., ISBN 80-214-2702-7
2. HOMIŠIN, J. et al.: *Základy strojného inžinierstva*. Košice: TU-SjF, 2001, 253 s., ISBN 80-7099-661-7.
3. KLIMO, V. – MANTIČ, M.: Some Problems on Micromechanism′s Actuators. In: *Acta Mechanica Slovaca*., roč. 8, č. 1, 2004, s. 69-74.
4. PAŠKO, J. – PAVLENKO, S. – KISLÍKOVÁ, I.: *Pružnosť a pevnosť 2: Teória a príklady.* Prešov: Fakulta výrobných technológií TU v Košiciach so sídlom v Prešove, 2003, 154 s., ISBN 80-7099-983-7
5. VOJTKO, I. – MATIJA, R.: Meranie a zhodnocovanie vibrácií. In: *Nové smery vo výrobných technológiách 2006: 8. medzinárodná vedecká konferencia, Prešov, 22. – 23. 6. 2006: Zborník referátov,* Košice: FVT TU, 2006, s. 503 – 508, ISBN 80-8073-554-9
6. BATEŠKOVÁ, Eva - MAŠČENIK, Jozef - NOVÁKOVÁ, Martina: Imagination support in depicting incisions and sections by means of PC. In: MOSIS '09, Czech republic. Ostrava: MARQ, 2009. p. 203-206. ISBN 978-80-86840-45-1.
7. Firemná literatúra fa. The MathWorks Inc. – MATLAB, Simulink, Simscape

Intelligent Plasma Monitoring Systems Based on Optical Methods and Chromatic Modulation

Dimitrios Tomtsis,
Technical and Educational Institute of West Macedonia,
Department of Applied Information Systems in Business and Economy,
Grevena, 51100, Greece

***Abstract*—In this work an intelligent optical monitoring system is presented the operating principle of which is based on natural biological and optical systems. The monitoring system perceives its environment thought optical sensors for goal directed actions even in dynamic changing environments and exhibits unsupervised and autonomous operation. Chromatic modulation methods are applied extensively to provide information compression in a form suitable for conventional image processing techniques. Chromaticity changes in a number of chromatic parameters are related to changes in physical plasma characteristics and properties (e.g. gas composition). An extensive range of parameters can be monitored with the same system so leading to the realization of a unified intelligent measurement philosophy.**

I. Introduction

Intelligent systems use ideas and get inspiration from natural systems and build on both established and novel techniques from a number of scientific fields such as computer vision, machine learning and artificial intelligence. Intelligent systems may perceive their environment through sensors for goal-directed actions even in dynamic situations while learning and functioning autonomously in order to cope with changing environments or inaccurate a-priori knowledge. Many intelligent optical systems incorporate those characteristics but are highly application dependent. Some systems are stand-alone applications which solve a specific measurement or detection problem, while others constitute a sub-system of a larger design. The specific implementation of an intelligent optical system also depends on if its functionality is pre-specified or if some part of it can be learned or modified during operation.

In many intelligent optical monitoring systems intensity modulation offers the advantages of inherent simplicity. However, conventional methods involving absolute intensity have associated problems. The most basic intensity monitoring systems use only a single photodiode to produce an output but these systems tend to be sensitive to spurious changes in intensity resulting from variations in the light source or other components within the system. Removing these spurious effects is difficult and leads to complicated and expensive systems.

Wavelength monitoring systems attempt to deduce the state of a system by taking the ratio of intensities at two different wavelengths. In principle, the two wavelengths should be chosen to be close enough together to ensure that spurious signals affect both wavelengths equally. However, the modulator needs to affect only one of the wavelengths so leading to conflicting demands. Such wavelength modulating systems may be constructed using spectrometers or by using two narrowband optical filters in conjunction with two photodetectors. Systems which make use of a spectrometer are expensive, optically inefficient and require excessive data processing. Systems using filters and photodetectors are wasteful of optical power.

Many of the difficulties inherent in the spectral or two wavelength monitoring methods may be overcome using a chromatic modulation method. A number of sensor systems based upon this approach have been developed [1] and shown to possess attractive advantages.

Taking into account the above considerations, it is desirable to develop a general purpose intelligent monitoring system the operation of which is not limited to a single task or application but it is reusable. In this work an intelligent optical monitoring system is presented which can be used in many different applications. The systems operation is based on chromatic modulation.

II. Biological Chromatic Processing

An aspect of the science of photic fields relates to the interaction of the photic field power $P(\lambda)$, which in general may be a complex function of wavelength λ with matter of different kinds, via its responsivity $R(\lambda)$, to produce a physical or chemical effect V [2] according to:

$$V = \alpha_1 \left[\int_{\lambda} P(\lambda) M(\lambda) R(\lambda) T(\lambda) d\lambda \right]^m \quad (1)$$

Where $\alpha 1$ is essentially a constant of proportionality, $T(\lambda)$ is the transmissivity of the propagating medium and $M(\lambda)$ is an externally imposed modulation. Eq. 1 describes the function of many diverse natural systems such as the production of chlorophyll by photosynthesis, radiation effects on animals and animal vision. In the later case, more detailed visual information is obtained via biological receptors having different responsivities $R1(\lambda)$, $R2(\lambda)$, $R3(\lambda)$.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_63, © Springer Science+Business Media B.V. 2010

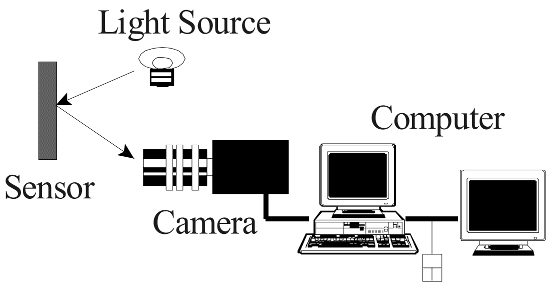


Fig. 1. Graphical illustration of the configuration of the system.

The significance of the above considerations for intelligent optical monitoring systems is that (1) also describes the detection of optical signals when R(λ) is the responsivity of the photodetector, V is the output voltage of the processing electronics, T(λ) is the transmissivity of the propagating system (such as an optical fiber) and α_1 is a constant of proportionality.

In general, chromatic changes can be monitored by a number (n) of detectors with overlapping spectral responses. The output of each detector may then be expressed as [3]:

$$V_n = \int P(\lambda) R_n(\lambda)\, d\lambda \tag{2}$$

where P(λ) is the spectral power distribution in the optical signal and $R_n(\lambda)$ is the wavelength responsivity of the n^{th} detector and λ is the wavelength. The color model representing this mathematical formalism is generally called RGB and it is widely used in self-luminous display technologies [4].

Each detector output may also be intensity normalised according to:

$$u_n = V_n / \sum_n V_T \tag{3}$$

and chromaticity maps may be formed in terms of the coordinates u_1, u_2, … $u_{(n-1)}$ (u_n becomes redundant since $\sum_n u_T = 1$).

The special case of n=3 leads to a two-dimensional chromaticity map (u_1:u_2) on which changes in optical signals may be traced. The special case when $R_1(\lambda)$, $R_2(\lambda)$, $R_3(\lambda)$ correspond to the responsivities of the human eye leads to the chromaticity map reducing to the CIE diagram of color science [5], [6].

The color model representing this mathematical formalism is called L_{XY} and provides the relative magnitudes of the tristimulus values (i.e. $x=u_1$; $y=u_2$; $z=u_3$).

Although the above method of displaying chromatic information has proved the information compression capabilities of the chromatic approach, it does not easily lead to the identification of signal changes in terms of fundamental signal properties. An alternative approach to the processing of the signals from such chromatic detectors overcomes this limitation. For the tristimulus case (n=3) this approach utilizes three chromatic parameters, namely dominant wavelength (H), intensity (L) and degree of monochromaticity or spectral width (S) which are defined as follows:

$$H = 120\left[m_i + \left(\frac{V_i - V_{\min}}{V_i + V_j - 2V_{\min}} \right) \right] \tag{4}$$

a) with $m_i=0$ for i=1, j=2, $V_{min}=V_3$, b) with $m_i=1$ for i=2, j=3, $V_{min}=V_1$, c) with $m_i=2$ for i=3, j=1, $V_{min}=V_2$

$$L = 100\left[\frac{V_{\max} + V_{\min}}{2} \right] \tag{5}$$

$$S = 100\left[\frac{V_{\max} - V_{\min}}{200\, m_2 - m_3 (V_{\max} + V_{\min})} \right] \tag{6}$$

with $m_2=0$, $m_3=-1$ for L≤50 and $m_2=m_3=1$ for L>50. V_{min} and V_{max} are the minimum and maximum detector outputs.

For the special case of $R_n(\lambda)$ corresponding to the responsivities of the human eye, H, L and S become the Hue, Lightness and Saturation of color science [7]. The color model representing this mathematical formalism is generally called HLS [8].

III. EXPERIMENTAL SYSTEM

The optical monitoring system is computer based and is broken down into three distinct and important elements: a camera, a color frame grabber card (CFG), and a host personal computer.

The camera converts the spatially distributed optical information into analog electric signals which are accepted by the CFG. The signal is then digitised and stored in the card's frame memory. The C.F.G. card is the interface between the camera and the host personal computer which is used to access and program the card's registers, frame memory, and buffers. Fig. 1 illustrates graphically the configuration of the system.

Although many measurands may be fully specified by distimulus detection [9] and dominant wavelength alone, others require tri-stimulus detection [10] to completely and unambiguously determine their state. The suggested intelligent optical monitoring system is tri-stimulus in nature and its hardware and software components are fully described in the following sections. The built-in intelligent nature of the monitoring system results from its ability to automatically compensate for any lightness variation as well as its ability to decide the gamma correction process based on the surrounding environment and the event being monitored.

A. Optical Frequency Response

The optical frequency response of the CCD camera was measured using the optical monitoring system and the apparatus shown in fig. 2.

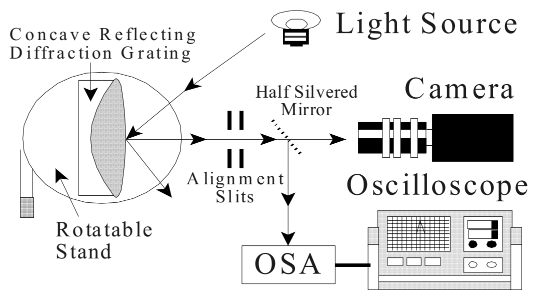


Fig. 2. Apparatus used for the optical frequency response of the color CCD camera.

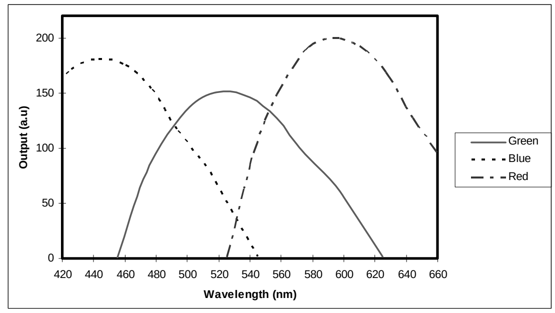


Fig. 3. Optical frequency response curves of the Red, Green and Blue camera detectors (tri-stimulus detection).

The required part of the spectrum was selected with the alignment slits and was directed to both the camera and the optical spectrum analyser (OSA) using the half silvered mirror. The oscilloscope was used to display information about the light intensity as a function of the selected wavelength incident to the camera and the OSA. Different wavelengths were selected by rotating the diffraction grating on the rotating stand. The resulting information was then processed by the host personal computer to provide color quantification that correlated with the wavelength monitored by the OSA.

The CCD array response of the color CCD camera was measured in a dark environment to avoid the unwanted effects of external lighting. The camera was used with its focus and aperture lens removed so that the light reflected from the grating was incident directly on the CCD element. Measurements were taken at 25 nm intervals.

The calibration results of the camera are shown in fig. 3. The obtained frequency response curves show that the CCD sensor is composed of three detector elements (tri-stimulus detection), with peak responses at 595, 525 and 445 nm for the red, green and blue detectors respectively, and spectral bandwidth 170 nm approximately for each detector.

The frequency response curve of each detector is utilised by eq. 1 and corresponds to the $R_n(\lambda)$ wavelength responsivity parameter.

B. Chromatic Calibration

The chromatic information received by the intelligent monitoring system depends on the lighting conditions of the surrounding environment. This means that the chromatic co-ordinates of an object under daylight conditions will be different than those under artificial light conditions. Other parameters that affect the chromatic measurements are lighting fluctuations, secondary light sources and shadowing. In a chromatic based system, this ambiguity which is involved with the measurement of an objects chromaticity makes it difficult to relate the optical emissions received from an object to the measurand. The RGB, Lxy and HLS color models are designed to be used with applications where the lighting conditions are not too different from daylight. Under different lighting conditions alternative procedures are usually adopted and incorporated into these models to allow for lightness adaptation.

The procedure adopted in this work is similar to the Von Kries transformation [11] with the difference being that instead of using a white color (or greys of the same chromaticity) as a reference, any carefully selected color can be used for the same purpose. In this procedure if a stimulus gives rise to the r,g, and b color responses then the resulting responses will depend on r/r_c, g/g_c, b/b_c, where r_c, g_c, and b_c are the chromatic coordinates of the reference color.

For a stimulus in the state considered to have the same color appearance as the same stimulus in a reference state it is necessary that:

$$r/r_c = r'/r_c', g/g_c = g'/g_c', b/b_c = b'/b_c' \tag{7}$$

where r_c', g_c' and b_c' are the color responses of the reference color in the reference state. Hence

$$r' = (r_c'/r_c)r, g' = (g_c'/g_c)g, b' = (b_c'/b_c)b \tag{8}$$

The resulting adapted variables r', g', b', obtained from (8), can then be used to calculate the L_{XY} tristimulus values as well as their HLS representation.

Hence, a calibration method for correcting the variations in RGB chromatic values was developed which concentrated on removing the RGB errors introduced by non-uniform illumination and external light fluctuations of the external environment.

The RGB values, the chromatic coordinates X, Y and the corresponding HLS values of a custom color patch were recorded under controlled laboratory lighting conditions [12], which have been used as reference information for any further processing.

C. Gamma compensation

An important characteristic of the human visual system is that it is sensitive to ratios of intensity levels rather than to absolute values of intensity. Therefore, the intensity levels should be placed logarithmically rather than linearly to achieve equal steps in brightness [13]. Gamma compensation is a facility available on most CCD cameras and it is applied to the R, G, and B detector responses to simulate the non-linear

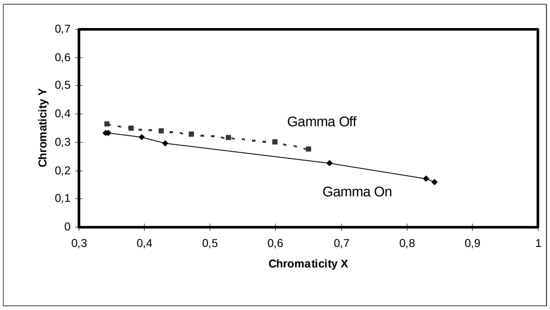


Fig. 4. Graph showing the resulting chromatic variation when gamma compensation is activated (full curve) and deactivated dashed curve.

function of the pupil of the human eye to provide high amplification of low intensity light and reduced gain for high intensity light and hence to provide larger intensity bandwidth for a given aperture setting. The effect of gamma compensation was investigated using the optical monitoring system of fig. 1, but in this case the sensor was replaced with a white card to provide a source of diffused white light. The X and Y chromatic coordinates were then evaluated for different aperture settings with the gamma compensation deactivated. This procedure was then repeated with the gamma compensation activated. The results are shown in fig. 4.

It is evident from this chromaticity graph that the chromatic changes due to gamma compensation are increased from {X,Y}={0.65, 0.27} to {X,Y}={0.85, 0.17}. This means that this facility can be used as a preprocessing tool in order to increase the contrast around an objects boundary and hence to enhance an algorithmic process.

IV. PLASMA MONITORING

In general, electric arcs consist of three physically distinct regions which correspond to two electrode regions and an interconnecting plasma column. The relative importance of the three regions in governing the overall behavior of the arc depends upon a number of factors which include the length of the arc gap, the type of arc and the electric power dissipated in the discharge.

Electric arcs may be subdivided into two main categories which are axis-symmetric and non axis-symmetric arcs. The axis-symmetric arc column burns symmetrically along the inter-electrode axis so providing good cylindrical symmetry, which is ideal not only for theoretical modeling but also for experimentation with sophisticated diagnostic techniques. A form of axis-symmetric arc without any boundary constraints is the '*free burning arc*' and is representative of furnace type arcs. Non axis-symmetric arc columns are less amenable to diagnosis because of the reduced cylindrical symmetry they provide.

Monitoring electric arc plasma is a difficult process because of the complexity of the phenomena which govern arc behavior. The physics of these processes and properties of the electric arcs are not fully understood due to the powerful

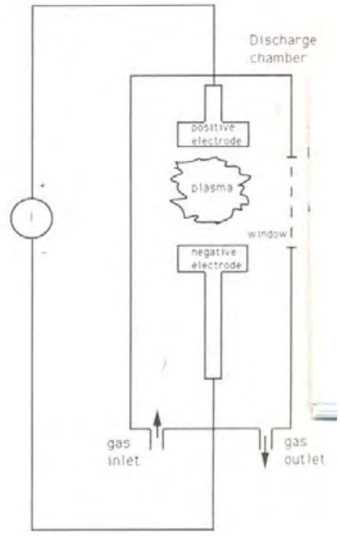


Fig. 5. Experimental apparatus for monitoring the electrical plasma.

mathematical analysis and considerable experimental effort required to model their complex characteristics [14]. For example, the arc plasma column has properties which vary with time and with radial, axial, and azimuthal coordinates. Phenomena at the arc electrodes are also complex and interactions with the plasma column as well as the interaction of the arc with its environment add to the complications [15].

Investigation of the electric arc phenomena has led to the identification of independent conditions which are simplified by eliminating some of the complexities described above, and perform detailed, highly accurate and localized measurements on that condition [16]. However, these measurements are made at the expense of determining how the overall electric arc and its environment behave and respond.

Spectroscopic detection and analysis techniques were also used in plasma monitoring applications but they are cumbersome and slow for on-line process control.

Often, such detailed knowledge is unnecessary for such process applications and all that is required is the identification of particular spectral signatures which are associated with certain quantities or parameters.

A. *Experimental Results*

Operational conditions in the plasma of such discharges need to be kept within specified tolerances and these conditions may be indicated by the spectral composition of the emission from the discharges.

The experimental arrangement for producing the plasma is shown in fig. 5. The discharge was DC in nature being maintained by a high voltage supply across the discharge electrodes. The optical monitoring system of fig. 1 was monitoring directly the discharge through a window at the side of the apparatus. A white light source was not used since the

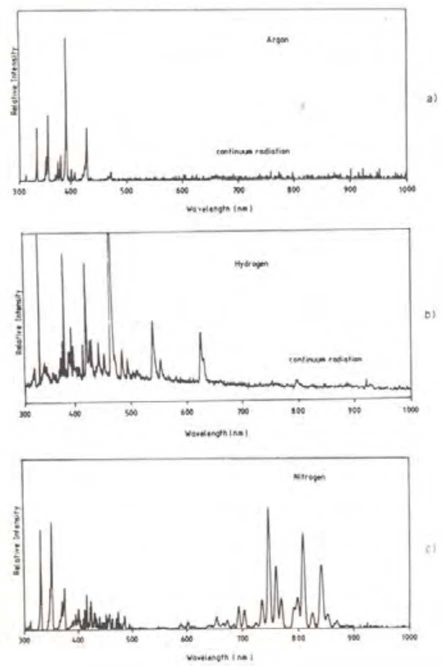


Fig. 6. Spectra of gas discharges used for plasma monitoring and processing. (a) Top: Argon plasma, (b) Middle: Hydrogen plasma, (c) Bottom: Nitrogen plasma.

arcing event by itself produced the necessary illumination. Neutral density filters were used to reduce the illumination of the plasma and to ensure that the operation of the camera was in the correct range by avoiding saturation of the CCD elements, and also to provide optical information concerning emission from the weakly emitting peripheral regions of the plasma column.

Both spectroscopic and chromatic measurements were taken and the emission from three commonly used gases for plasma processing (argon, hydrogen and nitrogen) investigated. The spectral signatures for the three gases are shown in fig. 6. The argon discharge is characterized by a

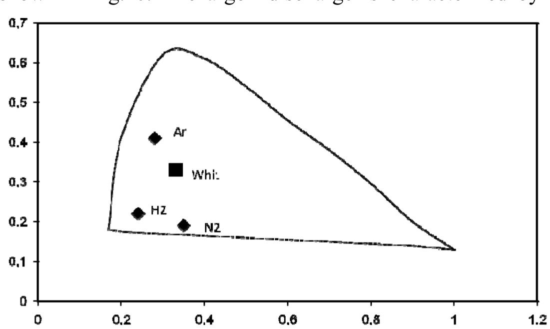


Fig. 7. Location on a chromaticity diagram of the emission from the plasma of Argon, Hydrogen and Nitrogen gases (diamond) and the white light point (square).

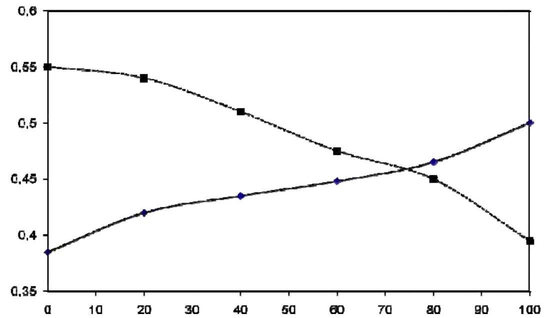


Fig. 8. Variation in the dominant wavelength (diamond) and intensity (square) of a plasma as a function of hydrogen concentration (x-axis).

number of spectral lines in the wavelength range 300-475 nm, hydrogen from 300-650 nm and nitrogen from 300-500 nm and 620-900 nm. For hydrogen and argon there is continuous emission apparent at the longer wavelengths.

The chromaticity values measured with the tri-stimulus system of fig. 1 are shown on the chromaticity diagram of fig. 7. The dominant wavelengths are 820, 920 and 875 nm for the argon, hydrogen and nitrogen respectively. The horse-shoe shaped chromatic boundary is also shown on the same figure. A set of experimental results demonstrating the effect on dominant wavelength and on intensity of changing the composition of the gases forming a DC plasma are presented in fig. 8. The results show the possible utilization of the chromatic approach for discriminating correct and incorrect process conditions.

V. Conclusion

The chromatic approach offers considerable scope for intelligent parameter monitoring with a wide range of principles and different optical sensing elements. Hence, there is the possibility of realizing an intelligent unified and reusable monitoring system whereby an extensive range of parameters are monitored using common forms of optical detection and signal processing methods in combination with various sensing elements.

References

[1] D. Tomtsis, V. Kodogiannis, "Optical pH Measurement Using Chromatic Modulation", International Conference of Computational Methods in Sciences and Engineering, ICCMSE 2003 Kastoria, Greece 12-16 September 2003.

[2] P. Moon and D.E. Spencer, "The Photic Field", Cambridge, MA:MIT, 1981.

[3] D. Tomtsis, V. Kodogiannis, "Digital Image Processing of Axis-Symmetric Electric Arc Plasma Based on Chromatic Modulation Methods", 6th WSES /IEEE Multi-conference CSCC-MCP-MCME 2002, Crete, July 2002.

[4] D. Tomtsis, K. Sapalidis, "A CCD-Based Tristimulus Colorimeter for Fast and Improved Measurement of Chromaticity Coordinates Of Displays by Matrix Correction Methods", WSEAS Transactions on Information Science and Applications, Issue 6, Volume 1, December 2004, pp. 1606-1611, ISSN: 1790-0832.

[5] CIE, Colorimetry Publication 15, 1971.

[6] CIE, Colorimetry Publication 15.2, 1986a

[7] P.C. Russell, J. Cosgrave, D. Tomtsis, A. Vourdas, L. Stergioulas and G.R. Jones, "Extraction of Information from Acoustic Vibration Signals Using Gabor Transform Type Devices", Measurement Science and Technology, Vol. 9, August 1998, pp. 1282-1290.

[8] J.D. Foley, A.V. Dam, S.K. Feiner, J.F. Hughes JF, "Computer graphics: Principles and practice", 1995, Addison-Wesley.

[9] D. Tomtsis, V. Kodogiannis, P. Chountas, "Distimulus Chromatic Measurement Systems", WSEAS Transactions on Circuits and Systems, Vol. 3, No. 2, April 2004, pp. 211-214, 2004. ISSN: 1109-2734.

[10] D. Tomtsis, K. Sapalidis, C. Katsanos, "Electronic Tristimulus Chromatic Measurement Systems", WSEAS Transactions on Circuits and Systems, Issue 9, Volume 3, November 2004, pp. 1835-1840, ISSN 1109-2734.

[11] R.W. Hunt, "Measuring Color", J. Wiley and Sons, New York, 1987.

[12] C. S. McCamy, H. Marcus and J.G. Davidson, "A color-rendition chart", Soc. Photog. Scientists Engineers, vol. 2, no 3, 1976, pp. 95-99.

[13] J.D. Foley, A. Van Dam, K.S. Feiner and J.F. Hughes, "Computer Graphics: Principles and Practice", Addison-Wesley, 1996.

[14] G.R. Jones, "High Pressure Arcs in Industrial Devices: Diagnostic and Monitoring Techniques", Cambridge University Press, Cambridge, 1988.

[15] G.R. Jones and M.T.C. Fang, "The Physics of High Power Arcs", Rep. Prog. Phys., 43, pp. 1415-1465, 1980.

[16] C.H. Flurscheim, "Power Circuit Breaker Theory and Design", Peter Peregrinus Ltd., Herts, 1975.

A Neoteric Chaotic & Adaptive Watermarking Scheme for Colored Images

Navita Srivastava* & Achintya Singhal†

*Department of Computer Science
Awadesh Pratap Singh University, Rewa
E-mail: navita.srivastava@gmail.com

†Department of Computer Science
Banaras Hindu University, Varanasi
E-mail: achintya.singhal@gmail.com

Abstract- *Today users are enjoying the convenience and advantages of advancement in technology. At present technology provides high quality multimedia processing, acquisition, flexibility, and reliability at significantly lower costs and ease. Digital multimedia documents authentication is thus one of the most important and investigated security applications. Watermarking schemes for the purpose has become evergreen research area. This paper outlines the previous work carried before by the authors and then proposes an image adaptive chaotic algorithm for watermarking of digital mages. The algorithm is tested against both image and text watermarks and found resilient to compression and filtering.*

Keywords: watermarking; data hiding; image authentication; multimedia security; chaotic, henon map.

I. Introduction

Quality work entails a lot of hard work, coupled with talent and use of intricate expensive equipments. But, the declining technological barriers, as a means to protect intellectual property, have emerged as the greatest threat. It is therefore obligatory to provide copyright protection for such work and retain confidence of brainpower. Cryptography tools available provide digital content integrity verification bit by bit but multimedia content authentication requirement is much different from it where by content enhancement/ worsening is allowed, without compromising the semantic content. This resulted use of watermarking techniques to hide the authentication information just inside the digital content because watermark becomes the part of the multimedia data and can be recovered to prove authenticity but can't be removed from it. The vital condition for watermarking is that the embedding technique should be such that it should produce the same results (multimedia document) even after the embedding.

Most watermark embedding processes are performed in either spatial domain or transform domain. In spatial domain watermarking schemes [24], [25], the watermark image is directly embedded into the host image by changing its pixel values. Both the insertion and extraction processes are relatively simple compared to transform domain watermarking schemes. However, it is more difficult for spatial domain watermarks to achieve imperceptibility because of the need to embed high intensity digital watermark for robustness tends to degrade the image visual quality. In transform domain watermarking schemes [6], [7], [8], [9], [11], [13], [15], [17], and [20], transform domain coefficients of the host image are modulated by the watermark information hence makes highly difficult to extract the watermark for the attacker and also achieve high quality watermarked image.

This paper uses an image oriented chaotic factor for hiding both text and colored logo watermarks in the targeted multimedia data as per algorithm outlined in [5]. The effective size of the image watermark is significantly reduced before embedding. The proposed algorithm extracts a graceful watermark from the authenticated image which is resilient to high quality JPEG compression, image format transformations like PNG, TIFF etc and image enhancements like color correction and filtering like high pass, Gaussian noise etc. This is obtained by avoiding the lossless entropy coding step of common image compressing algorithms. The paper is organized into sections: section II describes the watermark embedding and extraction process followed by results and discussion in section III and conclusions in section IV.

II. The Process

A. The background

The watermark insertion process is done in transform domain using Discrete Cosine Transformation (DCT) where a reversible, linear transform action is used to map an image into a set of transformed coefficients, which are then quantized and coded. The watermark insertion strength is calculated chaotically depending on the host image block pixel value. Henon map is used.

For image processing, the definition of two-dimensional DCT, block-sized n×n, is given by [23]:

$$F(u,v)=\sum_{j=0}^{n-1}\sum_{k=0}^{n-1}\left[\begin{array}{c} f(j,k)\propto(u)\propto(v)\times \\ \cos\left[\frac{(2j+1)\quad u\pi}{2n}\right]\cos\left[\frac{(2k+1)\quad v\pi}{2n}\right]\end{array}\right]$$

$$\propto(u),\propto(v)=\begin{cases}\sqrt{\frac{1}{n}} & u,v = 0\\ \sqrt{\frac{2}{n}} & u,v = 1,2,\ldots,n-1\end{cases}$$

for u, v = 0,1, … , n;

Here each transformed coefficient in an n×n DCT block is denoted as F(u,v).

The original image, F and watermark, W is DCT transformed; F is watermarked by selecting suitable image DCT blocks and incorporating the DCT coefficients of W with DCT coefficient of F, leaving behind the DC value [10]. DCT transformation provides resistance against JPEG compression and since it is based on HVS, the embedding in the domain helps in achieving imperceptibility.

The watermark embedding strength is decided chaotically for the selected block using Henon map. Chaotic systems are those whose state evolves with time – that may exhibit dynamics that are highly sensitive to initial conditions. Even though chaotic systems are deterministic dynamical systems, but sensitivity to initial conditions results the behavior of

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_64,

chaotic systems appears to be random. In general, the chaotic model system is given as:

$$x(n) = f(x(n-1))$$

where $x(n)$ is a chaotic sequence generated by the nonlinear map $f(\cdot), x(0)$ is the initial condition.

We propose the use of model introduced by French astronomer Michel Hénon called Hénon Map [21]. The Hénon map is a discrete-time dynamical system and takes a point (x, y) in the plane and maps it to a new point

$$x_{n+1} = y_n + 1 - ax_n^2$$

$$y_{n+1} = bx_n$$

The map depends on two parameters, **a** and **b.** For values of **a** and **b** the map may be chaotic, intermittent, or converge to a periodic orbit. But, generally, for smaller iterations, the model has chaotic behavior. The figure 1 shows the Henon plot generated for a particular image block. Figure 2 shows the incorporation strength factor generated for each pixel, to be used for watermarking in the block (8 X 8).

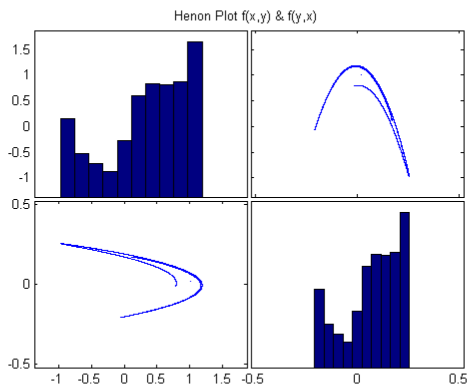


Figure 1: Henon plot generated for a block and corresponding histogram in x-y and y-x direction

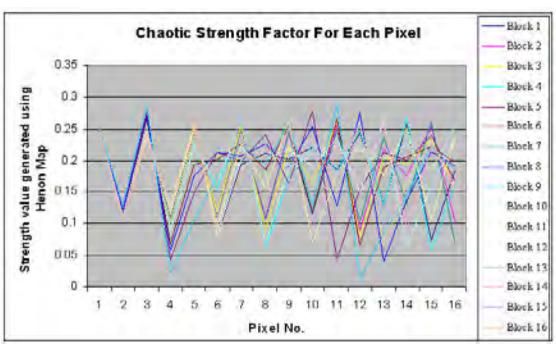


Figure 2: Showing chaotic strength factor for each pixel

The blocks suitable for watermarking and shedding the featureless blocks is computed using two techniques in conjunction i.e. edge detection and quantization matrix. Since an image is 2-dimensional representation of an object capturing its characteristics against some environment or background, it can be segregated into areas representing important parts and areas representing flat and featureless regions.

Edge detection algorithm mark the points in the digital image at which the luminous intensity changes sharply. Sharp changes in image properties usually reflect important events and changes in properties of the world. Hence, generating an edge map of the image and breaking it into equi-sized blocks (8x8), gives blocks with high activity blocks as discussed in [1], [2], and [3]. Also, the image is quantized using DCT and JPEG quantization matrix. This helps in identifying blocks resistant to JPEG compression and also high activity blocks. Common blocks from the two processes are then selected.

B. The Embedding Process

The graphical representation of the embedding process is given in figure 3. The original image and the watermark is DCT transformed and the blocks suitable for watermarking are segregated as discussed above. For each suitable block, watermark embedding strength factor is calculated using Hénon map. The watermark to be incorporated is also DCT transformed. Incase of image watermarks, the watermark is lossy compressed and doubly DCT transformed before embedding. With the calculated strength factor and the provided watermark, the watermark is embedded as discussed in [2].

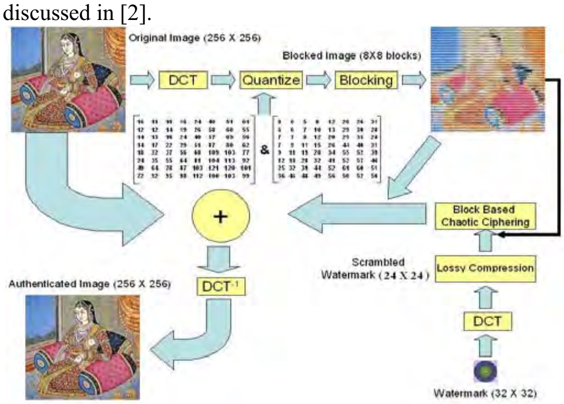


Figure 3: Watermark embedding process

C. The Extraction Process

The graphical representation of the extraction process is given in figure 4. This is a reserve process of embedding and requires original image and the image to be authenticated. The watermarked image will yield incorporated watermark else no significant result can be computed.

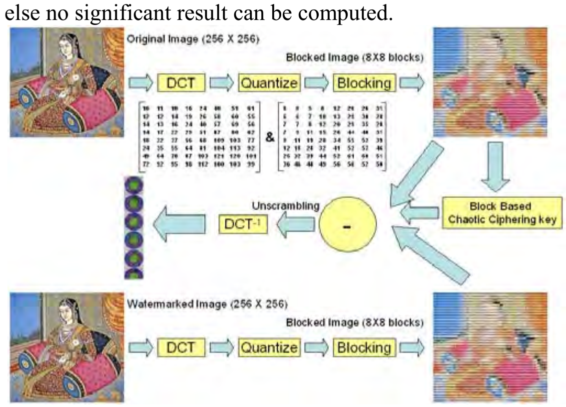


Figure 4: Watermark extraction process

III. Results

The results show the original image (figure 5, 13) along with its edge map (figure 6), the watermarked image (figure 7, 14), and the watermark used. The proposed scheme is tested with both the colored watermark (figure 8) and the text watermark (figure 15). The multiple copies of extracted watermarks (figure 9-12 and figure 16-19) are shown.

The calculation of NC [19], Q [26], PSNR and PSNR-HVS [18] gives the image quality index. The value of PSNR above 35dB is considered to very good. The PSNR value below 10dB is not good and below 5dB is unacceptable. The values of NC and Q lies in the range of 0 and 1 and of PSNR-HVS lie in the range of 0-100db. Higher the value, better are the results. Computation of NC, Q, PSNR and PSNR-HVS of extracted watermark is shown.

.

Results with colored image as watermark

Figure 5:Original Colored Image (240x240)

Figure 6:Edge Map of the image (240x240)

Figure 7:Watermarked Image NC=0.9999, Q=0.9980, PSNR=37.458db, PSNR-HVS=86.656db

Figure 8: Watermark Used

Figure 9: Watermark extracted from the original watermarked image

Extracted Watermark	1	2	3	4	5	6
NC	0.92089	0.9911	0.95733	0.99972	0.98248	0.9999
Q	0.78094	0.97751	0.65726	0.97407	0.94143	0.97479
PSNR	17.453	27.632	14.466	26.846	23.490	26.917
PSNR-HVS	62.223	76.584	58.958	74.708	69.103	74.488

Figure 10: Watermark extracted JPEG compressed (100% quality) watermarked image

Extracted Watermark	1	2	3	4	5	6
NC	0.9210	0.99323	0.96215	0.9997	0.97752	0.99652
Q	0.74085	0.96754	0.65941	0.95702	0.91934	0.95946
PSNR	16.578	26.03	14.503	24.617	22.035	24.827
PSNR-HVS	61.252	73.772	59.024	71.071	67.342	71.357

Figure 11: Watermark extracted from JPEG compressed (70% quality) watermarked image

Extracted Watermark	1	2	3	4	5	6
NC	0.85414	0.99265	0.94215	0.99247	0.98831	0.99541
Q	0.51453	0.8469	0.49458	0.84481	0.82711	0.86554
PSNR	13.167	18.965	11.979	18.858	18.269	19.050
PSNR-HVS	57.889	64.311	56.7	64.124	63.304	64.278

Figure 12: Watermark extracted from JPEG compressed (25% quality) watermarked image

Extracted Watermark	1	2	3	4	5	6
NC	0.96133	0.88003	0.90673	0.93168	0.90451	0.9888
Q	0.38554	0.27369	0.26757	0.35738	0.31649	0.46923
PSNR	9.8404	9.6085	8.6569	10.037	9.5879	10.812
PSNR-HVS	54.496	54.391	53.578	55.044	54.305	55.757

Image Quality Values with JPEG Compression								
Image Quality	**Image**				**Extracted Watermark***			
	Q	**NC**	**PSNR**	**PSNR HVS**	**Q**	**NC**	**PSNR**	**PSNR HVS**
25	0.97287	0.99578	32.8936	82.3866	0.37312	0.92453	9.7836	54.5583
30	0.97543	0.99608	33.2570	82.9433	0.42885	0.93820	10.5597	55.4103
35	0.97776	0.99652	33.6040	83.4483	0.45301	0.93008	11.1291	55.9846
40	0.98031	0.99675	34.0080	83.9406	0.54114	0.96950	12.0780	56.9800
45	0.98325	0.99720	34.5336	84.5243	0.60567	0.98016	12.9420	57.8166
50	0.98595	0.99762	35.1153	85.1850	0.71419	0.99179	14.8720	59.8380
55	0.98868	0.99858	36.1213	88.5916	0.73065	0.98468	15.4103	60.2820
60	0.98971	0.99874	36.4823	89.2483	0.76950	0.98489	16.2626	61.1966
65	0.99133	0.99899	37.1650	90.6863	0.80732	0.98413	17.3220	62.1320
70	0.99282	0.99927	37.9546	92.2746	0.86967	0.99053	19.0036	63.9820
75	0.99397	0.99948	38.7236	93.8706	0.92029	0.99159	21.4696	66.5353
80	0.99530	0.99973	39.9446	95.9230	0.95746	0.99892	24.2180	69.3803
85	0.99586	0.99989	40.5943	97.9793	0.96939	0.99954	25.8053	71.0510
90	0.99687	0.99997	42.3833	98.4450	0.97823	0.99980	27.3986	72.5283
95	0.99749	0.99999	44.4423	99.5000	0.98510	0.99956	29.2146	74.5810
100	0.99782	1.00000	46.5123	99.9970	0.98599	0.99990	29.2693	74.3253

* Extracted watermark quality measure is calculated against the watermark recovered from original watermarked image.

Results with text as watermark

Figure 13: Original Colored Image (240x240)

Figure 14: Watermarked Image NC=0.9999, Q=0.99379, PSNR=32.579db, PSNR-HVS=82.632db

THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT, AWADHESH PRATAP SINGH UNIVERSITY REWA MP

Figure 15: **Text Watermark Used**


THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUuER sCIENCE DEPARTMENT,
DWADIESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIaNCE DEPÄRTMaNT,
AWAÄHESh PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL &$NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRAPARARILFI EFITUZ

Figure 16: Watermark extracted from the original watermarked image


THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIaNCE DEPÄRTMeNT,
AWAÄHESh PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL &□NAVITA sRIVÄSTAVA, cOMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRATAP SINGH UNIVERSITY REWA MP
THIS WORK BELONGS TO ACHINTYA SINGHAL & NAVITA SRIVASTAVA, COMPUTER SCIENCE DEPARTMENT,
AWADHESH PRAPARARILFI EFITUZ

Figure 17: Watermark extracted from JPEG compressed (100% quality) watermarked image

Figure 18: Watermark extracted from JPEG compressed (91% quality) watermarked image

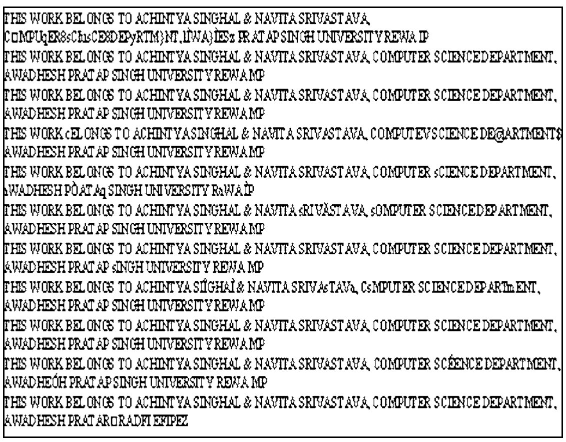


Figure 19: Watermark extracted from JPEG compressed (85% quality) watermarked image

The scheme, with colored watermark, shows excellent resistance to various attacks like JPEG compression up to 25% quality, color correction, brightness and contrast alterations, diffusion, high pass filter, uniform and Gaussian noise, rotations with 90 and 180 degrees etc. Using text as a watermark has got steganographic value more than digital watermarking. Its prime objective is to convey the hidden message to the receiver of the file. Its sustenance even with JPEG compression increases it application.

The proposed algorithm used meaningful colored watermark as well as text watermark in contrast to most of the previous work where either random number bit stream [14], or some known bit stream [16], or binary image [9]. The algorithm incorporated 8 pixels per selected block (8x8) for colored watermarks and 8 bits per selected block (8x8) for text watermarks which is significantly high in comparison to 3 bits per block (8x8) in [17], first three AC components of selected block (8x8) in [14] and 3840 pixels from 768x512 size image [16]. Further, the use of chaotic technique helped in attaining so called randomness in strength factor in comparison to fixed strength factor [9], [12], [14], and [16].

IV. Conclusion

In this proposed watermarking scheme, the strength factor is computed chaotically and dynamically for each block and each pixel wherever the insertions have to be made. Further, the watermark is embedded only in sensitive areas of the image thus securing the actual concerned content of the image. For identifying sensitive areas, image feature extraction using edge detection and JPEG quantization matrix is deployed. Moreover, the concept exploited only the Y component of YCbCr color space for watermarking as Cb and Cr components are more prone to loss during JPEG compression as our human visual system is dull for changes in these components [22]. The quality of extracted watermark (colored logo) from the unhampered image deteriorated in comparison to the algorithms proposed in [1], [2], [3], and [4] due to compression of the watermark used. But, the compression helped in achieving more insertions in lesser space thus retrieval in case of cropping is more promising.

The concept of double DCT introduced for colored watermarks to incorporate greater sustainability and provides cryptic sense to blind attacks being made. The algorithm proved to be generic and the results depict good sustenance to compression too with both image and text watermarks. The quality metrics of the extracted watermark depicted better results than previous algorithms [1], [2], [3], and [4]. Adobe Photoshop CS and Adobe ImageReady CS have been used for JPEG compression and various attacks.

The values of NC, Q, PSNR, PSNR-HVS reflect the imperceptibility of the watermark as well as the quality of retrieved watermark samples. The use of Hénon map guarantees that the watermarking strength for each pixel becomes inadvertent. Redundant inclusions and chaotic strength factor provides significant robustness and helps in obeying Kerckhoff's principle [27].

V. References

[1] A. Singhal, N. Srivastava and R. Mishra, "Edge sensitive frequency domain watermarking", in Proc. *International Information Technology Conference,* IITC, Colombo, Sri Lanka, pp. 127-134, Oct. 2006.

[2] A. Singhal, N. Srivastava and R. Mishra, "Hiding signature in colored image", in *Proc. IEEE International Symposium on Communication and Information Technologies,* ISCIT, Bangkok, Thailand, pp. 446-450, Oct. 2006.

[3] A. Singhal, N. Srivastava and R. Mishra, "Robust and sensitive watermarking scheme for Colored Images", in *Proc. 41st annual convention CSI*, Kolkata,, India, pp. 13-17, Nov. 2006.

[4] A. Singhal and N. Srivastava, "Digital Watermarking of Colored Images Using DCT-HSI Transformation", in *Proc. 40th annual convention CSI,* Hyderabad, India, pp. 201-204, Nov. 2005.

[5] A. Singhal and N. Srivastava, "Chaotic Watermarking Scheme for Colored Images", *The Icfai University Journal of Information Technology*, vol. 4, no. 4, pp. 23-33, Dec. 2008.

[6] C. H. Chang, M. Zhang, and Z. Ye, "A content-dependent robust and fragile watermarking scheme", in *Proc. 2nd International Conference on Visualization, Imaging and Image Processing*, pp. 201–206, Sep. 2002.

[7] C. H. Chang, Z. Ye, and M. Zhang, "Fuzzy-ART based digital watermarking scheme", in *Proc. IEEE Asia Pacific Conference on Circuits and Systems*, APCCAS, vol. 1, pp. 425–426, Dec. 2002.

[8] C. Rey and J. L. Dujelay, "Blind detection of malicious alterations on still images using robust watermarks", in *Proc. IEE Secure Images and Image Authentication Colloquium*, London, U.K., pp. 7/1–7/6, Apr. 2000.

[9] C. T. Hsu and J. L. Wu, "Hidden digital watermarks in images", *IEEE Trans. Image Process.*, vol. 8, no. 1, pp. 58–68, Jan. 1999.

[10] Chun-Shien Lu, "Multimedia security: steganography and digital watermarking Techniques for Protection of Intellectual Property", *Idea Group Publishing*, 2004.

[11] E. Izquierdo, "Using invariant image features for synchronization in spread spectrum image watermarking", *EURASIP Journal of Applied Signal Processing*, vol. 4, pp. 410–417, 2002.

[12] H. Inoue, A. Miyazaki, and T. Katsura, "Wavelet-based watermarking for tamper proofing of still images", *International Conference on Image Processing Proceedings, ICIP 2000*, pp. 88-91, Oct 2000.

[13] J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia", *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[14] J. Huang, and Y. Q. Shi. "Adaptive image watermarking scheme based on visual masking", *Electronics Letters,* vol. 34, no. 8, pp. 748-750, Apr 1998.

[15] J. J. Eggers and J. K. Su, "A blind watermarking scheme based on structured codebooks", in *Proc. IEE Secure Images and Image Authentication Colloq.*, London, U.K., pp. 4/1–4/21, 2000.

[16] J. J. K. Ó. Ruanaidh, W. J. Dowling, and F. M. Boland, "Watermarking digital images for copyright protection", *IEE Proceedings on Vision, Signal and Image Processing*, vol. 143, no. 4, pp. 250-256, Aug 1996.

[17] J. Zhao and E. Koch, "Embedding robust labels into images for copyright protection", in *Proc. Int. Congress IPR for Specialized Information, Knowledge and New Technologies*, Vienna, Austria, pp. 242–251, Aug. 1995.

[18] K. Egiazarian., J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS", in *Proc. Second International Workshop on Video Processing and Quality Metrics, Scottsdale*, 4 p, 2006.

[19] M. A. Suhail and M. S. Obaidat, "A robust digital watermarking technique", *IEEE,* pp. 629–632, 2000.

[20] M. Barni, F. Bartolini, V. Cappellini, and A. Piva, "A DCT-domain system for robust image watermarking", *Signal Processing*, vol. 66, no. 3, pp. 357–372, 1998.

[21] M. Henon, "A two-dimensional mapping with a strange attractor", *Communications in Mathematical Physics,* vol. 50, pp. 69-77, 1976.

[22] M. Ejim and A. Miyazaki, "A wavelet-based watermarking for digital images and video", in *Proc. International Conference on Image Processing,* ICIP, vol. 3, pp. 678-681, Oct. 2000.

[23] R. C. Gonzalez and R. E. Woods, "Digital Image Processing", *Pearsons Higher Education 2/E*, 2002.

[24] R. G. Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark", in *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. 86–90, Nov. 1994.

[25] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding", *IBM System Journal*, vol. 35, no. 3&4, pp. 313–336, 1996.

[26] Z. Wang and A. C. Bovik, "A universal image quality index", *IEEE Signal Processing Letters, XX,* 2002.

[27] http://en.wikipedia.org/wiki/Kerckhoffs'_principle

Rotary Transducer

Ilie Borcosi[1], Nicolae Antonie[2], Alina Dinca[3]
"Constantin Brancusi" University
Calea Eroilor, no.30, Targu-Jiu, Romania
[1]ilie_b@utgjiu.ro, [2]antonienicolae@yahoo.com, [3]alina@utgjiu.ro

Abstract – **Using the speed voltage generators to measure the rotary of a driving system is an advantageous solution for the analogical structures of the command circuit. For the numerical systems we will need to use an analogical-digital converter and for this reason these will be more useful and economical using the numerical transducer. Trough this paper we propose a numerical rotary transducer which offers advantageous price/ performances report.**

The purpose of this application is to diagnose a process automatically, with no human analysis involved.

I. Introduction

The control of a driving system of physical process presumes the obtaining of information about it, information that need to be worked on, memorized or communicated to a higher level. The physical process is determined by physical values that can be transformed to electrical signals (usually analogically by nature) by the transducer obtaining in this way information about the process. We can work on this signals using analogical technique or numerical technique case in witch the analogical signals given by the transducers are converted into numerical signals using special electronic circuits named data acquisition systems.

The speed transducer or rotary transducer will give information about the speed of the motor during an application with the diving system. Usually we use a speed voltage generator of direct current to give the information about the necessary speed. Because this transducer has an analogical voltage at its output, proportional to the speed of the motor, we have to use an analogical – digital converter if we want to obtain the numerical value of this signal, used in a numerical system. If we know the constant of the speed voltage generator, kTG [V/rad/s], and the maximum value of the speed of the motor, Ωmax [rad/s], we can determine the maximum value of the tension at the output of the speed voltage generator which needs to be measured, $V_{Gmax} = k_G * \Omega max$ [V]. Usually, the connecting scheme of the output of the speed voltage generator to the entrance of the analogical – digital converter has to be adjusted (by using a simple resistive divisor) used to adapt the variation beach of this signal to the accepted values beach from the entrance of the analogical-digital converter (usually +/- 10 V or +/- 5 V). The converter resolution in the driving system applications is usually 10-12 bits.

The position transducer makes the measurement of the driving system position. An advantageous solution is offered by the utilization of the incremental optical encoder which eliminated the analog digital converters [1]. The encoder is formed of the following components: emitters (LED), receivers (photodetectors), rotating codewheel, stationary mask and the electronic circuit for the handling of the signals. The two wheels (rotating and stationary) are divided in the opaque and transparent zones. When the transparent zones of both wheels are superposed, the receivers get the light flux from the emitters and transmit an impulse to the output. It is a relative transducer that gives a number of specified impulses for one rotation of its ax, which depends on the number of the transparent zones. This impulse rotation number will lead to the transducer resolution that can be considered like an analogical-digital converter of the position.

For a number of N rotation impulses the equivalent parameters of the position transducer will be:

-the variation beach of the output: N (bits);

-the transducer resolution: [log2N] (bits);

-the precision of the transducer: $2\pi/N$ [rad/bit].

Because the encoder doesn't give information about the absolute position, we will use a special interface, basically a numerical circuit, to count the impulses given by the encoder. At the same time, because using only one train of impulses doesn't allow the obtaining of the rotation sense, the encoder give 2 trains of rectangular impulses at 90 degrees one against the other. A sense detection circuit can be designed so that it could, based on the notification of this difference phase (advance or delay with 90 degrees of one signal against the other) to determine incrementing and decrementing of the position info.

Because this transducer is relative, which doesn't point to the position modification facing the starting position of the system, the encoders give another numerical signal, the also called zero signal of the encoder. This signal becomes active giving a pulse by the length of an encoder step just one time on the rotation and it can be also used for the initialization after putting it under supply voltage so we can set an absolute position info of the system.

In the figure 1 we can see the shape of the signals given by the encoder (A and B the two measurement canals; Z the initial position canal) [2]. It is possible the using of each one of the positive and negative edge of the signals A and B of the encoder which allows the increasing by 4 times the precision of the encoder by implementing some adequate decoding circuits.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_65, © Springer Science+Business Media B.V. 2010

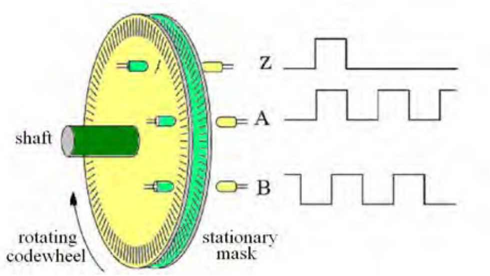


Fig.1. The signals generated by the encoder to output.

II. SYSTEM DESCRIPTION

A. Presenting the system

In the figure 2 is presented a structure for the revolution transducer designed and made by us. It is formed from a disk (figure 2.a), placed on the motors ax, disk on which is made an orifice Or [3]. This orifice allows the passing of the bright flux, from a transmitter to a receiver set face to face on a scale in the figure 2.b. These are sorted in a way so that the orifice Or to pass through each group emitter – receiver.

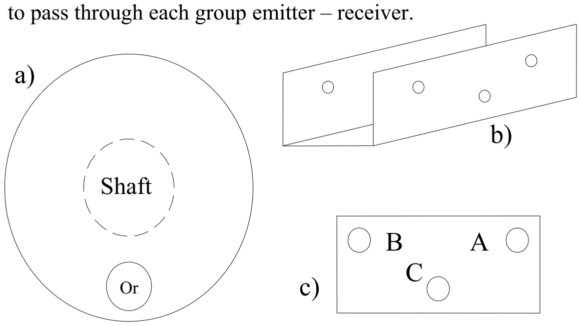


Fig. 2. The element component of the transducer

We use infrared sensors so that the natural light doesn't influence the work. The three sensors A, B, C (figure 2.c) transmit at the output V_{osen}=1 logical when they are illuminated. The signal from the output of the sensors (V_{osen}) is applied to a handling circuit (derivation) made with Schmitt Trigger inverters (figure 3) to eliminate the errors which appear in the cases of slow fronts (at small rotations). In this way, at the illumination of every receiver, the handling circuit generates a short impulse (the signal V_C from the figure 3 which correspond to the C sensor).

The A and B sensors are used to determine the rotation sense, and the C sensor which obtains the value of the rotation which sets up a counter which increments the tact impulses.

The rotary will be determined using the formula:

$$n = (f_i/ T_c)*60 = 1/(T_i*T_c)*60 \text{ [circ/min]},$$

where fi=1/Ti represents the frequency of the impulses tact applied to the counter and T_c the value of the counter (number of increment impulses which are equivalent to a temporization – Temp).

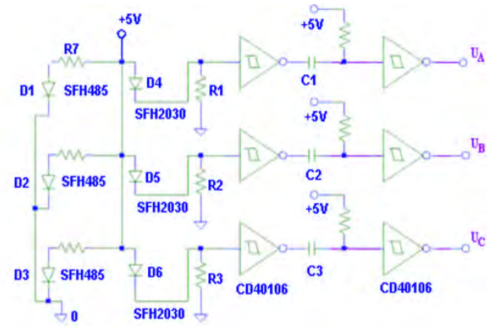


Fig. 3. The circuit for handling the sensors signals A, B, C.

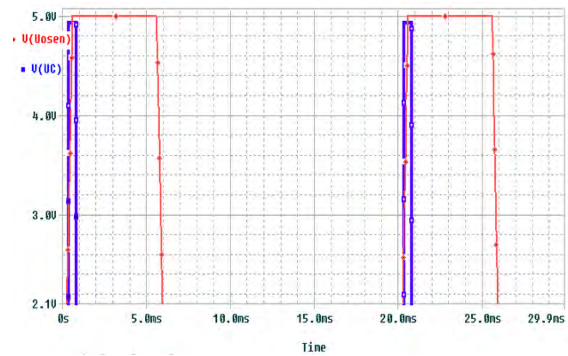


Fig. 4. The input signals (Vosen) and output (Vc) from the handling circuit.

B. System needs

The determination of the sense and the value of rotation are made by implementing the logical diagram from figure 5. According to the logical diagram, first we attribute to the rotation n and the variable v_1 value zero, after that we check the A sensor. If A=0 we verify C, and if C=0 then we verify B. In case that A=1 logical and C=1 logical we set the counter (which will start to implement the impulses with f_i frequency) and the control variable v_1 (which is used to approve the evidence after a complete rotation). If follows B=1 and then C=1 we made a complete rotation (v_1=1) and because of these succession of activating the sensors we can tell that the transducer disk makes a rotation to the right, we calculate the value of the revolution n (we process the value to the Temp counter). After that we reset the contain of the counter and prepare for a new rotation. If the succession of illumination of the sensors is C=1, A=1, B=1 and then C=1 it means that the transducer makes a revolution to the left. When one of the successions of a rotation sense (left or right) is not kept, we reset the counter and verify which the new succession is. To verify the eventual stopping of the motor in the countering time, we inserted a procedure in the figure 6, where the current value can be compared to the reference value T. This value (T) represents the maximum time interval as long as the counter can increment ($T_{cmax-preset}$=$Temp_{max-preset}$), between the activity of two sensors.

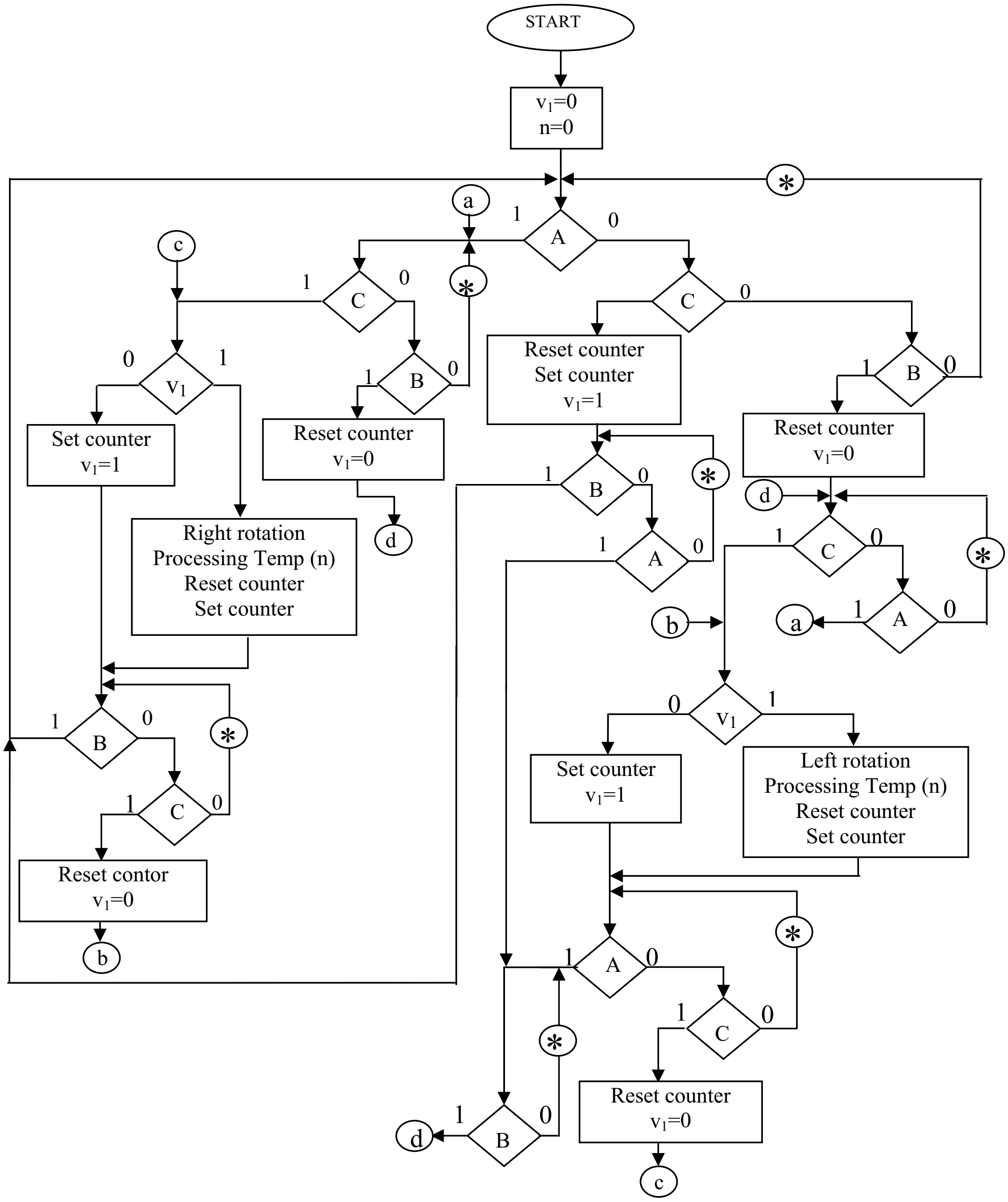


Fig. 5. The logical diagram.

It is chosen so that we can measure the rotation $n \geq n_{min\text{-}preset}$, where $n_{min\text{-}preset}$ is better to have a small value, resulting for $T_{cmax\text{-}preset}$= $Temp_{max\text{-}preset}$ a bigger value. We consider that the motor is stopped if Temp>T ($n < n_{min\text{-}preset}$) when the counter is reset and it's attributed the variable v_1 and the rotation n has zero value.

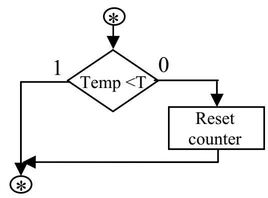


Fig.6. The verification procedure for stopping the motor.

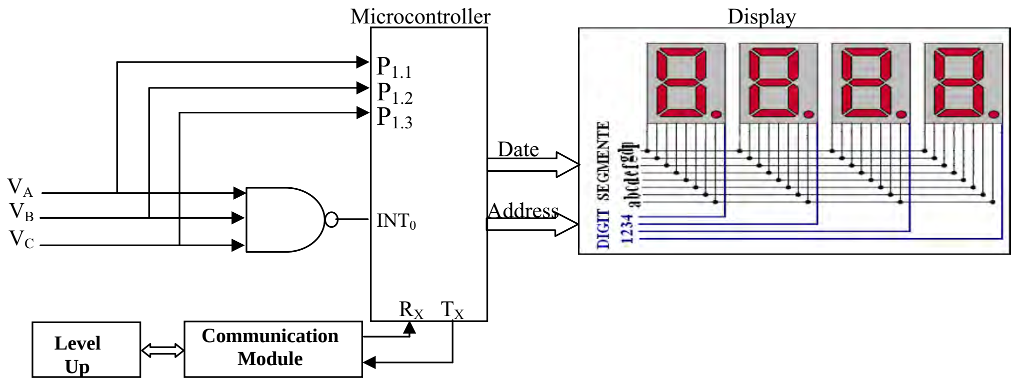


Fig.7. The processing circuits with microcontroller.

C. Making the system

Handling the signals given by the circuit in fig. 3 (V_A, V_B, V_C) which shows the sensors' state, it can be made with the help of a microcontroller or the calculator through implementation of the logical diagram in fig.5, according to this example we used a microcontroller type AT89C2051 (according to the figure 7) [4]...[6].

The signals V_A, V_B, V_C are connected to the specific input to the external interruption INT0, and at the same time, one separately to another inputs P1.1, P1.2 and P1.3 which determine if the signal is activated. The connecting to the INT0 input should be done using a logical gate OR which adds the three signals, but, considering that the interruption becomes active on the negative edge then the signals must be negated and applying the principals of Boolean algebra in the place of OR gate and using the NAND gate. The timer/counter function is made by counting the impulses with a frequency equal to the oscillator frequency divided by 12. The temporization frequency choosing the oscillator frequency f_{osc}=12 MHz is f_i=1MHz and the period is T_i = 1µs. The timer/counter circuit is on 16 bits. We can temporize maximum: $2^{16}*T_i=65536*1\mu s=65536\mu s \approx 65ms$.

It is preferred to obtain a temporization of 50 ms, other wise if we want a bigger temporization which is easier to multiply. These value of temporization result if we load the temporization circuit with the initial value 15536, this temporization until the value 65536 and each time are touched these value a counter is incremented.

To detect the stop of the motor in the timer/counter time we considered $n_{min\ preset}$=0,1 rot/s, which means that if in 10 seconds a full rotation is not made, it means that the motor is stopped and the timer/counter circuit is cleared, resulting for T the value of 10 s. The temporization value is 10 seconds and is obtained by measuring 20 times 50 ms.

III. Conclusions

The measurement system of rotation designed and made by us has the following advantages:

-it's easy and simple to make and offers advantageous report price/performances because it has the easier processing digital system (a single microcontroller) than optical incremental encoder [1];

-it can be used for monitoring rotation by keeping the dates in a data base or by showing it;

-it does need an A/D converter, in the case of a numerical control system, can be cuter the handling circuit with a microcontroller, its function will be implemented in the control and command system. Actually we acknowledge a trend to realize the numerical exclusive systems with alone microcontroller witch provides the all control and command functions;

-on the disk we can make a lot of orifices. The real rotation will be equal to the calculated rotation multiplied to the orifices number. Is preferred that the number of orifices to be a multiple of 2, and the distance between them must not be bigger than the distance between the A and B sensors;

- we used this rotation transducer and we recommended it to be used in drive systems with the DC motor controlled by a computer (using the parallel port) which has a single external interrupt input. To this input we connect more signals for real time control, for example, synchronization input (for control in phase) and speed transducer outputs. The proposed speed transducer will enable rarely the interrupt input (for each of its signals, once at a complete rotation), compared with an encoder, at which the activation depends on the number of pulses.

References

[1] Fl.Grofu, L. Popescu, *Digital system for optical incremental encoder connection,* U.P.B. Sci. Bull., Series c, vol. 69, no. 3, 2007.
[2] Rotary Encoder, http://hades.mech.northwestern.edu/wiki/index.php./Rotary_Encoder.
[3] O. Olaru, I. Borcosi, M.C. Popescu, O. Tulpan, N. Antonie, *The revolution measurement of a d.c. machine*,The XIIth Edition of the International Scientific Conference of the Engineering Faculty, November 23rd-24th 2007, Tg.-Jiu.
[4] L. Toma, *Sisteme de achizitie si prelucrare numerica a datelor,* Ed. de Vest, Timisoara, 1996.
[5] D. Marinescu, *Microcontrolerul 80C32 Manual de utilizare,* Editura Tehnica, 1998.
[6] AT89C2051 and AT89C52 datasheet, available at: http://www.atmel.com.
[7] Fl. Grofu, L. Popescu, C. Cercel, *Acquisition System for Monitoring Vibration,,* CISSE 07 - IETA 07, University of Bridgeport, USA, December 03-12, 2007, Published by Springer Press.
[8] I. Borcoşi, O. Olaru, M.C. Popescu, A. Dinca, N. Antonie, M. Ionescu, *Device with Analogical Circuits for Protection to the Lack of the Pulse for the Three-Phase Rectifiers in Electrical Drive*, INTERNATIONAL JOURNAL NAUN, Issue 4, Volume 2, December 2008.

Regression Analysis about Humidity Elimination from Diesel Fuel Via Bioorganic Compounds to Increase Antifouling Action

C.G. Tsanaktsidis[1], N. Sariannidis[2] and S.G. Christidis[1]
[1]Laboratory of Qualitative Fuel Control
[1]Department of Pollution Control and Technologies/ [2]Department of Financial Applications
Technological Education Institute of Western Macedonia,
Kila, Kozani –Greece 50100

Abstract: **In this paper we try to export a regression equation which describes the variation of humidity of Diesel fuel with time. A hydrophilic polymer is used to eliminate humidity from diesel fuel and the regression analysis shows that the maximum decrease of humidity in an hour it is of about ~ 35%.**
Keywords: *Regression Analysis, Diesel Fuel, Humidity Elimination Bioorganic Compounds, Hydrophilic Polymer Antifouling Action.*

I. INTRODUCTION

The quality control of alternative liquid fuel aims at finding out if fuel is appropriate so that the decrease in the pollution of the environment can be secured, as well as the right function of combustion engine and the maximum output.
One of the most important properties of fuels is the water presence, which is undesirable because it causes corrosion, can cause ice blockage in the power grid that can be formed in low temperatures and last it can be divided during the warming-up and cause flame extinguishing from the produced steam (water vapour).
In this paper it is attempted to describe the modification of humidity of diesel fuel via an equation. For this aim we introduce in different samples of diesel fuel the hydrophilic polymer TPA (which has the property to retain water molecules) and its synthesis procedure presented by a suite of software, ChemOffice.
The values of humidity is recorded in the modification table by time and with these results is exported the regression equation [1-4] which describes the variation.
It is stressed that the volume of the fuel and the mass of the polymer maintain stable.

II. EXPERIMENTAL AND STATISTICAL PROCESS

A. *Synthesis of Hydrophilic Polymer, Thermal Polyaspartate Anion (TPA)* [9,10,11,16]

The synthesis of polymer is held through the condensation of D,L aspartic acid (1.5g) towards the equivalent polysuccinimide [12] and the hydrolysis of the latter towards the final product of this reaction, the thermal polyaspartate anion (TPA).

The selection of aspartic acid as a raw material was supported by the variation of humidity degree that these particle groups show in a water environment.[15]

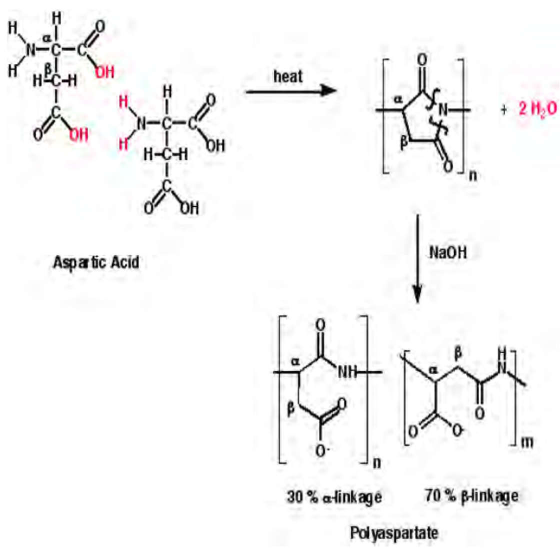


Fig 1. Synthesis of Hydrophilic Polymer, Thermal Polyaspartate Anion (TPA).

B. *Humidity Elimination of Diesel Fuel via TPA.*

At first the humidity of Diesel fuels is determined with the method ASTM D-1744[13] via an automatic potentiometer titrator Karl-Fisher.
Next, 0.1 g of TPA is introduced in samples of Diesel 20 mL volume each one. Then, TPA is eliminated through filtration.[14]
The values of humidity are recorded every 30 minutes in Table I.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_66, © Springer Science+Business Media B.V. 2010

Volume of Diesel (mL) + Mass of TPA (g)	Time (t)	Humidity (mg/g)
20 + 0	0	0.0650
20 + 0.1	30	0.0635
20 + 0.1	60	0.0489
20 + 0.1	90	0.0520
20 + 0.1	120	0.0600
20 + 0.1	150	0.0478
20 + 0.1	180	0.0492
20 + 0.1	210	0.0455
20 + 0.1	240	0.0630

Table I. Humidity Values by Time of Diesel Fuels with TPA.

It must be referred that 20 mL is the demanded quantity of Diesel fuel for the determination of humidity according to method ASTM D-1744, while the quantity 0.1 g of polymer is defined from the limits of chemical additives in liquid fuels which is 1-4 Kg/tn.

C. Statistical Empirical Findings

The preliminary statistic analysis of sequences of humidity Y and time T (see Table I) and especially the relevant scatter plot have shown that the modifications of Y, probably determined by the cosine of T. Specifically, in this paper we have examined the regression model $Y_i=b_0+b_1Sin(T_i)+u_i$ in order to investigate the relation between humidity of Diesel fuel and the time, when the volume of the fuel and the mass of the polymer maintain stable.

Some diagnostic tests were performed to establish goodness of fit and appropriateness of the model. First, it was examined whether the standardized residuals and squared standardized residuals of the estimated model are free from serial correlation. The results show (see Table II) that the LB statistics for the standardized residuals and standardized squared residuals are not significant [8]. In addition, the independence of the standardized residuals is confirmed by the Durbin-Watson statistics 1.44 [7]. Also, in the Table IV the ARCH–LM Test concerning two lags in the residuals ($N*R^2 = 1.55$) showed that the variance does not exhibit heteroskedasticity [6] that has been confirmed by the application of White test ($N*R^2 = 2.27$). Finally, with the application of the Jarque-Bera test (1.27) it was found that the normality of the residuals u_i cannot be rejected [5].

Table IV presents results of the regression. The adjusted R^2 (0.37) indicates that modifications of the humidity of fuel (Y) affected considerably by the Time (T) that the hydrophilic polymer TPA remains at the fuel. Specifically, the positive coefficient b_1 (0.00667) of cosine time and its statistical significance is interpreted as the fuel's humidity is decreased periodically with time when the fuel's volume and the polymer's mass that is introduced in fuel are constant. The periodicity of the phenomenon via the max-min of humidity values shows that the maximum decrease of humidity in an hour it is of about ~ 35%.

Standardized Residuals				Squared Standardized Residuals			
Lags	Auto Correl - ation	Partial Correl- ation	LB(n)	Lags	Auto Correl -ation	Partial Correl- ation	LB(n)
1	0.193	0.193	0.4629	1	-0.286	-0.286	1.0149
2	0.048	0.011	0.4951	2	-0.28	-0.394	2.1206
3	-0.295	-0.318	1.9330	3	-0.086	-0.400	2.2427
4	-0.376	-0.301	4.7266	4	0.263	-0.083	3.6170
5	-0.217	-0.104	5.8903	5	-0.117	-0.249	3.9540
6	-0.074	-0.095	6.0702	6	-0.021	-0.151	3.9683
7	0.147	-0.006	7.1333	12	0.031	-0.096	4.0153
8	0.074	-0.158	7.6720	24	-0.005	-0.172	4.0176

Table **II**. LB Test for Residuals and Squared Residuals from the Regression Estimation.

Note: *LB(n) are the n-lag Ljung-Box statistics for the residual series. LB(n) follows chi-square variable with n degree of freedom; the series of residual contains 9 observations .*

Squared Residuals lag(-1)	Squared Residuals lag(-2)	F-statistic
-1.533816	-0.40	0.573
(-0.84)	(-0.87)	

Table III**.** ARCH–LM Test for Squared Residuals.

Notes: *Figures in parentheses are t-statistics.*

b_0	b_1	Adjusted R^2
0.055*	0.006667**	0.37
(0.001983)	(0.002795)	

Table IV**.** Mean Equations $Y_i=b_0+b_1Sin(T_i)+u_i$.

Notes: *Standards errors are shown in parentheses.*

**indicates statistical significance at the 1% level.*

***indicates statistical significance at the 5% level.*

III. CONCLUSIONS

The study investigates how the changes of the fuel's humidity were influenced by the Time (T) that the hydrophilic polymer TPA remains at the fuel using a regression model. More precisely, it was found that the change of humidity should be described via regression equation at which the dependent variable is the changes of the humidity and the independent variable is the cosine of the time that the hydrophilic polymer TPA remains at the fuel.

The analysis of the regression estimation showed that: the decrease of humidity is changed periodically with time and comes to its maximum value at the first hour.
The use of equation can predict the development of humidity values for important time periods and gives the moment for the filtering of fuel so that the humidity is removed in the biggest degree. The elimination of humidity makes the fuel used in the car machines and gives combustion with less pollution for the environment.
It is necessary to study the description via equations of the phenomenon associates with volume of fuel, mass of polymer in order to export a total equation.
The result of elimination of humidity from fuel gives the possibility to be rendered as antipollution at his use, so the analysis of economic cost of the whole process is required when it is applied in the industrial production.
Collaboration of sectors of science, organic chemistry, technology of fuels, statistical analysis for the export of conclusions in the qualitative control of fuels.

REFERENCES

[1] Chuvieco E.; Riaño D.; Aguado I.; Cocero D., *Estimation of fuel moisture content from multitemporal analysis of Landsat Thematic Mapper reflectance data: applications in fire danger assessmen*t, publisher Taylor and Francis Ltd, International Journal of Remote Sensing, Vol. 23, No 11, **2002** , pp. 2145-2162(18).

[2] MINGHOU XU*, YAOGUO FAN, JIANWEI YUAN, CHANGDONG SHENG AND HONG YAO, *A SIMPLIFIED FUEL-NOx MODEL BASED ON REGRESSION ANALYSIS, INTERNATIONAL JOURNAL OF ENERGY RESEARCH*, Int. J. Energy Res., 23, 157Ð168 (1999)

[3] David Carter, Daniel A. Rogers, Betty J. Simkins, *Does Fuel Hedging Make Economic Sense? The Case of the US Airline Industry,* AFA **2004** San diego Meetings, September 16, **2002**.

[4] Gerfried Cebrat, Avraam Karagiannidis and Agis Papadopoulos, *Proposing intelligent alternative propulsion concepts contributing to higher CO_2 savings with first generation biofuels*, Management of Environmental Quality: An International Journal, Vol. 19 No. 6, **2008**, pp. 740-749

[5] Jarque C. M, Bera A. K**. (1987)** *A test of normality of observations and regression residuals.* International Statistical Review 55: 163-172

[6] Engle R. F. Autoregressive conditional (1982) *Heteroscedasticity with estimates of the variance of United Kingdom inflation*. Econometrica 50 (4): 1987-1007.

[7] Durbin J, Watson G.S. (1950) *Testing for serial correlation in least square regression.* Biometrika 37: 409-428

[8] Ljung G. M, Box G.E.P. (1978) *On a measure of lack of fit in time series models*. Biometrika 65: 297-303

[9] *The Presidential Green Chemistry Challenge Awards Program: Summary of 1996 Award Entries and Recipients*; EPA744-K-96-001; U.S. Government Printing Office: Washington, DC, 1996

[10] J. Clark and D. Macquarrie *Handbook of GREEN CHEMISTRY AND TECHNOLOGY* 2002

[11] Wheeler, A.P, Koskan L.P *"Large scale thermally synthesized polyaspartate as a substitute in polymer applications"* Mat. Res. Soc Symp. Proc, 1993, 292,277.

[12] J. Mosig, C.H Gooding, A.P Wheeler*, "Kinetic and thermal characterization on the hydrolysis of polysuccinimide"* Ind. Enf. Chem. Res. 1997. 36. 2163-2170

[13] ASTM D1744-92 *Standard Test Method for Determination of Water in Liquid Petroleum Products by Karl Fischer Reagent* (Withdrawn **2000**)

[14] ASTM D2276 - 06 Standard Test Method for Particulate Contaminant In Aviation Fuel by Line Sampling ,**2006**

[15] A.N.Troganis, C.Tsanaktsidis and I.P. Gerothanassis, *^{14}N NMR relaxation times of several protein amino acids in aqueous solution-comparison with ^{17}O NMR data and estimation of the relative hydration numbers in the cationic and zwitterionic forms, J. Magn.Reson.,* **2003**,164, 294

[16] S.G. Christidis, A.I Maroulis, K.Chatziantoniou -Marouli , A.Tsiomlektsis, Aristotle University of Thessaloniki. *«The experiment in teaching of Chemistry. Green Approach. Composition and Applications of Thermal Polyasparte Anion.»*. Proceedings of 20th Pan-Hellenic Conference of Chemistry. Ioannina September **2005.**

Optimal Threshold for Locating Targets Within a Surveillance Region Using a Binary Sensor Network

Zhenxing Luo and Thomas C. Jannett
Department of Electrical and Computer Engineering
The University of Alabama at Birmingham
Birmingham, AL 35294

***Abstract* - This paper considers the design of the optimal threshold for locating targets within a specified surveillance range using a network of binary sensors. The threshold was determined by minimizing a cost function representing the summation of the variances of estimation errors for the *x* and *y* coordinates of the target. The cost function was evaluated using multiple integration over the specified ranges for target power and target location. A geometrical interpretation of the optimal thresholds for targets with known power and location is presented to offer insight into the problem. The optimal threshold was validated in Monte Carlo simulations using a field of sensors having a uniform grid layout. In simulations, the summation of the variances of the estimation errors achieved using the optimal threshold approached the minimum of the cost function.**

I. Introduction

TARGET localization is an important and challenging problem in sensor networks [1]-[4]. In a network or network region, sensors transmit data to a central processing node or fusion center. For the target localization problem, the processing node receives data from each sensor and estimates the target location. Some methods are based on time difference of arrival (TDOA) [5], [6]. However these methods require accurate timing, which is not feasible with inexpensive sensors. Energy-based methods, which use signal strength information, are more practical in real situations [7]-[10]. Much recent work has been focused on maximum likelihood target localization in networks of inexpensive sensors that have limited energy and communications resources [1], [9]. When resources are limited, it is desirable for the sensors to transmit quantized data to the processing node [1], [11]. A maximum likelihood estimator that uses quantized data and the Cramer-Rao lower bound (CRLB) for this estimator are derived in [1]. The CRLB is the theoretical performance bound for the variances of the estimation errors achieved using the maximum likelihood estimator. The CRLB has been used as the basis for several methods used to control the network localization performance, including designing detection thresholds, evaluating network layout, and reconfiguring a network after loss of resources [1], [12].

In this paper, we extend the work in [1] by developing an optimal threshold for binary sensors detecting targets having a range of powers and locations within a specified surveillance region. The paper is organized as follows. Since this paper extends the work in [1], Section II closely follows the presentation and notation in [1]. Section II presents the sensor model, exemplary network layout, and maximum likelihood estimator for quantized data. Section II also reviews a method for designing optimal quantization thresholds for targets with known power and location. Section III provides a geometrical interpretation of the optimal thresholds for targets with known power and location. Section IV develops the optimal threshold for binary sensors locating targets having a range of powers and locations within a specified surveillance region. Methods for computing the optimal threshold are presented and demonstrated. The optimal threshold is validated and evaluated in Monte Carlo simulations. Discussion and conclusions are presented in Sections V and VI.

II. Problem Formulation

We consider an exemplary field of N = 441 sensors uniformly deployed with sensors placed 9 m apart in each of the x and y directions (Fig. 1). Many ideas presented in this paper apply for other sensor deployments, including fields in which sensor are placed at random locations.

The target emits a signal, with the signal intensity decreasing as the distance from the target increases. Following the presentation in [1], we consider a signal power model

$$a_i^2 = \frac{G_i P_0'}{\left(d_i / d_0\right)^n} \tag{1}$$

where a_i is the signal amplitude at the ith sensor, G_i is the gain of the ith sensor, and P_0' is the power emitted by the target measured at a reference distance d_0. The Euclidean distance between the target and the ith sensor is

$$d_i = \sqrt{(x_i - x_t)^2 + (y_i - y_t)^2} \tag{2}$$

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_67,

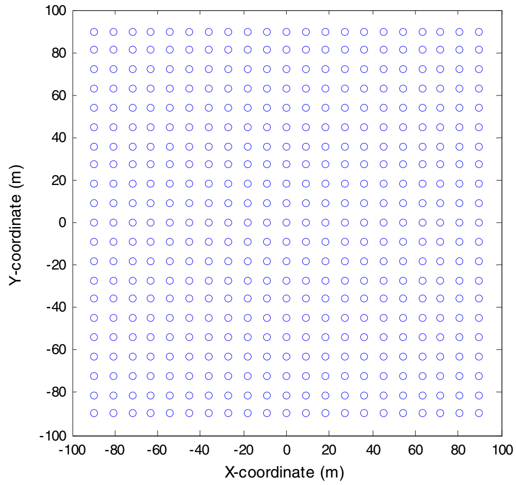


Fig. 1. Sensor field layout showing 441 sensors placed 9 m apart.

where (x_i, y_i) and (x_t, y_t) are the coordinates of sensor i and the target, respectively, and n is the power decay exponent. We assume that $G_i = G$ for $i = 1,...,N$, such that $P_0 = GP_0'$. In this paper, we use n = 2 and we assume that the target is at least d_0 = 1 meters away from any sensor at all times, resulting the simplified signal power model

$$a_i^2 = \frac{P_0}{d_i^2}. \quad (3)$$

The measured signal at the ith sensor is modeled as

$$s_i = a_i + w_i \quad (4)$$

for $i = 1,...,N$ where w_i is a Gaussian noise

$$w_i \sim N\left(0, \sigma^2\right). \quad (5)$$

At each sensor, the measured signal s_i is quantized and transmitted to the processing node.

As described in [1], the processing node receives quantized multi-bit data from each sensor. The received data are $D = \{D_i : i = 1,...,N\}$, where D_i can take on any discrete value from 0 to $2^M - 1$. Using the definition $L = 2^M$, the set of quantization thresholds for the ith sensor is $\vec{\eta}_i = [\eta_{i0}, \eta_{i1}, ..., \eta_{iL}]$, where $\eta_{i0} = -\infty$ and $\eta_{iL} = \infty$. The ith sensor supplies quantized data

$$D_i = \begin{cases} 0 & -\infty < s_i < \eta_{i1} \\ 1 & \eta_{i1} < s_i < \eta_{i2} \\ \vdots & \vdots \\ L-2 & \eta_{i(L-2)} < s_i < \eta_{i(L-1)} \\ L-1 & \eta_{i(L-1)} < s_i < \infty \end{cases}. \quad (6)$$

The probability that D_i takes specific value l is

$$p_{il}(\vec{\eta}_i, \theta) = Q(\frac{\eta_{il} - a_i}{\sigma}) - Q(\frac{\eta_{i(l+1)} - a_i}{\sigma}) \quad (0 \le l \le L-1) \quad (7)$$

where $Q(\cdot)$ is the complementary distribution function of the standard Gaussian distribution

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt. \quad (8)$$

After collecting the data, D, the processing node estimates the parameter vector $\theta = [P_0\ x_t\ y_t]^T$ by maximizing the log-likelihood function of D

$$\ln p(D|\theta) = \sum_{i=1}^{N} \sum_{l=0}^{L-1} \delta(D_i - l) ln\left[p_{il}(\vec{\eta}_i, \theta)\right] \quad (9)$$

where

$$\delta(x) = \begin{cases} 1, & x = 0 \\ 0, & x = 0 \end{cases}. \quad (10)$$

In summary, the maximum likelihood estimate of $\theta = [P_0\ x_t\ y_t]^T$ is determined through the optimization problem:

$$\max_\theta \ln p\left(D|\theta\right). \quad (11)$$

The maximum likelihood estimate (11) and the corresponding Fisher Information matrix (FIM) and CRLB for the estimation error variance are derived in [1], which also addresses the design of sensor decision thresholds based on the CRLB using both optimization and heuristic methods.

In this paper, we considered binary decisions made at each sensor that yield binary data

$$D_i = \begin{cases} 0 & -\infty < s_i < \eta \\ 1 & \eta < s_i < \infty \end{cases}. \quad (12)$$

A. Optimal Quantization Thresholds for Fixed Power and Fixed Target Location

The CRLB for the ML estimate (11) is a function of the quantization thresholds and was used in [1] as the basis for the cost function

$$V(\vec{\eta}) = \frac{J_{11}(\vec{\eta})[J_{33}(\vec{\eta}) + J_{22}(\vec{\eta})] - J_{13}^2(\vec{\eta}) - J_{12}^2(\vec{\eta})}{|J(\vec{\eta})|} \quad (13)$$

which represents the summation of the variances of the location estimation errors in the x and y directions. In (13), J_{ij} is the element on the ith row and jth column of the FIM, J. Expressions for J_{ij} are presented in [1] and will not be reproduced here. The cost (13) is an obvious choice for designing a threshold that provides high accuracy in the estimation of the target location, although the resulting estimate of the power, P_0, may not be accurate. For a known P_0 the optimal threshold for a target located at specific coordinates (x_t, y_t) is the threshold that is the problem solution

$$\min_{\vec{\eta}} V(\vec{\eta}). \quad (14)$$

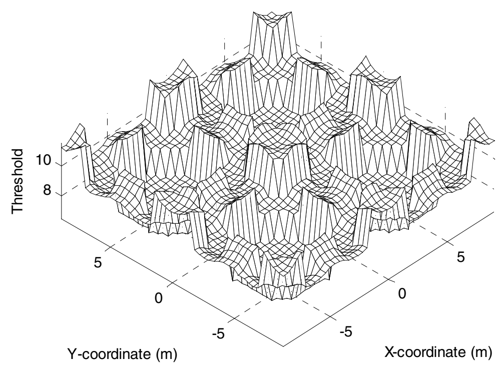
For a specified target location and power, this method gives the threshold that is optimal for use at all sensors in the field. However, the method does not address the practical problem of choosing the threshold needed to locate targets having unknown power and unknown positions within a surveillance region. Since the method for threshold optimization (14) depends on knowledge of the parameters $\theta = [P_0\ x_t\ y_t]^T$ to be estimated, different thresholds would be required for different target locations within a surveillance range of interest, and the thresholds are optimal only for a specific P_0. However, the parameters are not known ahead of time, so the method cannot be used in practice. Despite its practical limitations, the optimal threshold (14) serves as a valuable benchmark for evaluating thresholds developed using other methods [1].

In this paper, the optimal threshold (14) is studied further. In the next section, we present a geometrical interpretation of the patterns of sensors that fire at the optimal thresholds for demonstrative target locations. The geometrical interpretation provides insight into the threshold design problem and the limitations of the performance of the optimal threshold (14).

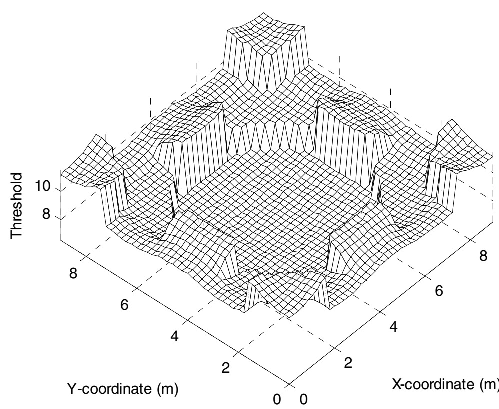
III. Geometrical Interpretation of the Optimal Thresholds for Fixed Power and Target Location

The optimal thresholds computed using (13) and (14) are valid only for a specified target location. In this section, we plot the optimal thresholds over a grid of target locations within a surveillance region and then examine the patterns of sensors that would fire if the optimal thresholds are applied.

For P_0 =10000 and binary decisions made at each sensor, optimal thresholds were determined over a subsection of the field (Fig. 1) represented by a 46 by 46 grid of target locations with x = 0.0, 0.2,…, 9.0 and y = 0.0, 0.2,…, 9.0. For this grid of target locations, the optimal threshold (14) varies symmetrically with target location from a minimum of 6.89 to a maximum of 11.24 (Fig. 2). For a target located far away from any sensor, the threshold is lower than for a target located near a sensor. The optimal threshold sometimes changes abruptly from one target location to another.



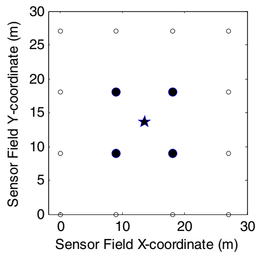
a) Optimal thresholds over a field subsection having vertices [-9, -9], [9, -9], [-9, 9], and [9, 9].



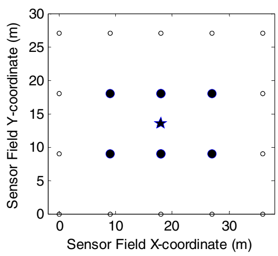
b) Detailed view of optimal thresholds over a smaller field subsection having vertices [0, 0], [0, 9], [9, 0], and [9, 9].

Fig. 2. Optimal thresholds computed using (14) for the corresponding target locations over subsections of the field of Fig. 1 (P_0=10000, binary decisions, σ= 1).

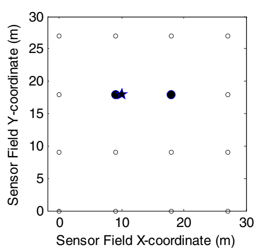
Next, we choose three demonstrative target locations for which the optimal thresholds used by sensors to generate binary decisions are at relatively low, medium, and high levels. We demonstrate the impact of the different thresholds by showing the patterns of sensors that fire (detect a target) for these thresholds at P_0=10000 (arbitrary units) in the noise free case (σ = 0). For a target located in the center of a grid of four sensors, the target is located as far away as possible from any of the four sensors, the threshold is low, and four sensors fire (Fig. 3a). For a target located an equal distance between two sensors, the optimal threshold is at an intermediate level and six sensors fire (Fig. 3b). For a target located close to a sensor, the threshold is relatively high and two sensors fire (Fig. 3c).



a) Target location (13.5,13.5) and optimal threshold 6.89.



b) Target location (18, 13.5) and optimal threshold 9.75.



c) Target location (10, 18) and optimal threshold 11.38.

Fig. 3. Demonstrative target locations (star) and corresponding patterns of fired sensors (solid circles) that occur for the optimal thresholds found using (14) if there is no measurement noise. For a), b) and c) the lower left sensor corresponds to location (0,0) in the field of Fig. 1. (P_0=10000, binary decisions, σ = 0).

Inspection of the sensor firing patterns for the optimal thresholds (Fig. 3) facilitates assessment of the impacts of noise and power mismatch. Noise may make more or fewer sensors fire, resulting in asymmetric sensor firing patterns that produce estimation errors. If the thresholds are the same over a range of target locations, then small changes in target location will not produce different firing patterns, resulting in poor identifiability (uniqueness of solution) and possibly, biased estimates. If the power is significantly lower than the power for which the threshold is optimal (power mismatch), few or no sensors may fire, resulting in large target location estimation errors [1].

Finally, we examine the consequences of applying suboptimal thresholds. If the applied threshold is higher than the optimal threshold, few or no sensors may fire, resulting in large estimation errors. If the applied threshold is lower than the optimal threshold, more sensors may fire than with the optimal threshold, perhaps with little consequence. For target locations near the border of the sensor field, a low threshold that results in a large pattern of sensors firing may produce an asymmetric firing pattern that introduces estimation errors. Further study of the log-likelihood function (9) and FIM or CRLB may provide insight into the relationship between performance, thresholds, and sensor firing patterns.

IV. Optimal Thresholds for a Range of Powers over a Surveillance Region

The minimization in (14) provides a threshold that is optimal only for a known P_0 and specified target location. However, our goal is to locate targets having unknown power and unknown positions within a surveillance region. In [1], the idea of integrating (13) over a surveillance region and range of powers was presented, but was considered impractical due to the prospect of a prohibitive computational load. In this paper, we use this idea for determining an optimal threshold for locate targets having unknown power and unknown positions within a surveillance region. First, we assume that the power is uniformly distributed within a specified interval, and we assume that the target locations of interest are uniformly distributed within a specified surveillance region. Next, we calculate an average cost function by employing multiple integration of (13) over the power interval and over the target locations within the surveillance region. In this approach, the elements of the FIM are understood to be a function of the threshold and the parameter vector $\theta = [P_0 \ x_t \ y_t]^T$, and the cost

$$\overline{V}(\vec{\eta},\theta) = \iiint \frac{J_{11}(\vec{\eta},\theta)[J_{33}(\vec{\eta},\theta) + J_{22}(\vec{\eta},\theta)] - J_{13}^2(\vec{\eta},\theta) - J_{12}^2(\vec{\eta},\theta)}{\left|J(\vec{\eta},\theta)\right|} dx_t dy_t dP_0 \tag{15}$$

is computed by integrating over θ.

The optimal threshold (14) designed for a specific power performs poorly for mismatched P_0 values [1]. When the actual power is lower than the value of P_0 used to calculate the optimal threshold, few or no sensors fire, and the location variance $V(\vec{\eta})$ computed using (13) becomes very high. Therefore, we expect that the cost (15) will be dominated by the contributions of the low target powers. In a uniform sensor field, for a fixed power, the optimal thresholds (Fig. 2) found using (14) for specific target locations and the corresponding values of $V(\vec{\eta})$ follow a symmetric pattern within the field. An understanding of the sensitivity of the optimal threshold to mismatched low P_0 values and recognition of the symmetry in $V(\vec{\eta})$ suggest that approximating the cost (15) as

$$\overline{V}(\vec{\eta},\theta) = \sum_{x_t}\sum_{y_t}\sum_{P_0} \frac{J_{11}(\vec{\eta},\theta)[J_{33}(\vec{\eta},\theta)+J_{22}(\vec{\eta},\theta)] - J_{13}^2(\vec{\eta},\theta) - J_{12}^2(\vec{\eta},\theta)}{\left|J(\vec{\eta},\theta)\right|} \quad (16)$$

may give useful results and avoid a prohibitive computational load in finding the optimal thresholds that are the solutions to

$$\min_{\eta} \overline{V}(\vec{\eta},\theta) . \quad (17)$$

A. Computation of the Optimal Thresholds

Optimal thresholds were computed for exemplary cases that demonstrate the use of (16) and (17). First, the optimal threshold was found for targets having a fixed power and locations within a surveillance region having vertices [-36,-36], [36,-36], [-36,36], and [36,36] in the field of Fig. 1. Second, the optimal threshold was determined for targets that have P_0 within a specified interval and locations within the surveillance region. Binary decisions were employed and $\sigma = 1$ in all cases.

The optimal threshold for a fixed $P_0 = 8000$ was determined for target locations within the surveillance region. The symmetry of the thresholds exhibited in Fig. 2 was exploited to approximately evaluate (16) for the surveillance region based on a grid of 81 target positions for x = 1,2,…9 and y = 1,2,…9. The minimization (17) was performed using exhaustive search. Optimal thresholds were also determined using the same method for $P_0 = 10000$ and for $P_0 = 12000$ for the same surveillance region (Table I).

TABLE I
OPTIMAL THRESHOLD AND COST VERSUS THE SET OF POWERS USED FOR OPTIMIZATION OVER A SURVEILLANCE REGION

Power, P_0	Optimal Threshold, η	Cost, $\overline{V}(\vec{\eta},\theta)$
8000	7.11	4.38
10000	7.91	4.05
12000	8.66	3.85
[9000 10000 11000]	7.80	4.14
[8000 10000 12000]	7.48	4.30

The optimal threshold for P_0 within the interval from 8000 to 12000 was determined for target locations within the surveillance region. The cost (16) was evaluated over a P_0 set [8000 10000 12000] by exploiting symmetry to approximately evaluate (16) for the surveillance region based on the grid of 81 target positions described earlier (Table I). The minimization (17) was carried out using exhaustive search. In consideration of Table II in [1], the cost was expected to be dominated by the contributions from the low powers within the P_0 interval. The cost was evaluated for P_0 within the interval from 8000 to 12000 using a P_0 set having only three elements because using more elements would be expected to increase the computational load without much improvement in the solution. In addition, for P_0 within the interval from 9000 to 11000, the cost (16) was minimized using the same method for a P_0 set [9000 10000 11000] (Table I).

B. Evaluation of Thresholds in Monte Carlo Simulations

The optimal thresholds were evaluated in Monte Carlo simulations (1000 runs, binary decisions, n=2, $\sigma = 1$) in which targets were placed at random locations uniformly distributed over the surveillance region represented by the square having vertices [-36,-36], [36, -36], [-36, 36], and [36, 36]. Simulations were initially performed for fixed P_0 values of 8000, 10000, and 12000 with the thresholds shown in Table II. Optimum thresholds matching the power or power intervals were studied. We also studied a threshold mismatch in which a suboptimal threshold was applied. The suboptimal threshold was the lowest of the optimal threshold values found for any target location within the surveillance region.

TABLE II
RESULTS OF MONTE CARLO SIMULATIONS EVALUATING THRESHOLDS AT DIFFERENT POWER LEVELS OVER A SURVEILLANCE REGION

Power, P_0	Optimal Threshold, η	Applied Threshold	Cost at Optimal Threshold, $\overline{V}(\vec{\eta},\theta)$	Sum of Location Estimation Error Variances Using Applied Threshold
8000	7.11	7.11	4.38	4.84
10000	7.91	7.91	4.05	4.21
12000	8.66	8.66	3.85	4.18
Random (8000-12000)	-	7.48[1]	-	4.92
10000	7.91	6.89[2]	4.05	4.73

[1] 7.48 is the optimal threshold found for P_0 set [8000 10000 12000] in (16).
[2] 6.89 is the lowest of the optimal threshold values found for any target location within the surveillance region (Fig. 2a).

Additional simulations were performed for random P_0 uniformly distributed within the interval 8000 to 12000. The sum of the experimental location estimation error variances for estimates of the target position (x_t, y_t) was computed and compared to the minimum cost (16) at the optimal threshold $\min_{\eta} \overline{V}(\vec{\eta},\theta)$ (Table II).

V. Discussion

The optimal thresholds determined for fixed P_0 levels of 8000, 1000, and 12000 increase with P_0 (Table I). The optimal threshold determined for the P_0 set [8000 10000 12000] was less than the threshold for the set midpoint P_0 = 10000, but was closer to the optimal threshold for the set midpoint P_0 = 10000 than to the optimal threshold for P_0 = 8000. The optimal threshold determined for the P_0 set [9000 10000 11000] was also less than the threshold for the set midpoint P_0 = 10000.

The costs at the optimal thresholds for fixed P_0 levels of 8000, 10000, and 12000 decrease as the power level used for optimization increases (Table I). The cost at the optimal threshold determined for the P_0 set [8000 10000 12000] was greater than the cost at the threshold for P_0 = 10000, and was closer to the cost for the optimal threshold for P_0 = 8000 than to the cost at the optimal threshold for P_0 = 10000. The cost at the optimal threshold determined for the P_0 set [9000 10000 11000] was greater than the cost at the threshold for P_0 = 10000. These results confirm that the costs, and corresponding optimal thresholds, were dominated by the contributions from the lower P_0 values.

Despite the approximations and the coarseness of the representation of the surveillance region and power interval used in evaluating (16) in the minimization used to find the optimal thresholds, the optimal thresholds gave good performance (Table II). In simulations, the sum of the error variances achieved using the optimal thresholds approached the minimum cost for all P_0 values considered. When a threshold mismatch, in which the applied threshold was not the optimal threshold, was studied for P_0 = 10000, the sum of the error variances achieved (Table II, line 5) slightly exceeded the sum of the error variances achieved using the optimal threshold (Table II, line 2).

VI. Conclusion

This paper addresses the important problem of decision thresholds for target localization. In this paper, the work of [1] was extended by developing and demonstrating a method for finding the optimal threshold used to make binary sensor decisions for targets having power and location within specified intervals. Symmetry was exploited to find the optimal thresholds using minimization performed over a subset of the surveillance region, which reduced the computational effort expended in finding the optimal thresholds. The optimal threshold resulted in target localization performance approaching the minimum cost. The new method for designing the optimal threshold for targets having power and location within specified intervals could be useful in designing a sensor field to meet target localization performance requirements.

A geometrical interpretation of the optimal thresholds that depend on knowledge of the target location and power was presented for a sensor field with sensors placed in a uniform grid pattern.

The optimal thresholds were found through minimization accomplished by approximate evaluation of (16) over a finite set of target locations (grid) within the surveillance region considered, and over a very coarse set of power values within the power interval considered. At the expense of a greater computational effort, better approximations and use of a finer grid might be employed to produce more accurate values for the optimal thresholds. Other future work could consider refinements in the methods used in computing (15)-(17), extension for multi-bit quantized data, detailed performance analysis, and random sensor layouts.

References

[1] R. X. Niu and P. K. Varshney, "Target location estimation in sensor networks with quantized data," *IEEE Trans. Signal Processing*, vol. 54, pp. 4519-4528, Dec. 2006.

[2] C. Cevher, F. D. Marco, and G. B. Richard, "Distributed target localization via spatial sparsity," *Proc.16th European Signal Processing Conference (EUSIPCO)*, Laussane, Switzerland, 2008.

[3] Y. Zou and K. Chakrabarty, "Sensor deployment and target localization in distributed sensor networks," *ACM Trans. Embed. Comput. Systems,* vol. 3, issue 1, pp. 61-91, Feb. 2004.

[4] K. Chakrabarty, S. S. Iyengar, H. Qi, and E. Cho, "Grid coverage for surveillance and target location in distributed sensor networks," *IEEE Trans. Computers*, vol.51, no.12, pp.1448-1453, Dec. 2002.

[5] K. H. Yang, G. Wang, and Z.Q. Luo, "Efficient convex relaxation methods for robust target localization by a sensor network using time differences of arrivals," *IEEE Trans. Signal Processing*, vol. 57, pp. 2775-2784, July 2009.

[6] E. Weinstein, "Optimal source localization and tracking from passive array measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.30, pp. 69-76, Feb. 1982.

[7] D. Li, K. D. Wong, Y.H.Hu, and A. N. Sayeed, "Detection, classification, and tracking of targets," *IEEE Signal Process. Mag.,* vol.19, no. 3, pp. 17-29, Mar. 2002.

[8] D. Li and Y. H. Hu, "Energy based collaborative source localization using acoustic microsensor array," *EURASIP J.Appl. Signal Process.,* no. 4, pp.321-337, 2003.

[9] X. Sheng and Y. H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. Signal Processing*, vol.53, no.1, pp. 44-53, Jan. 2005.

[10] N. Patwari and A. O. Hero, "Using proximity and quantized RSS for sensor localization in wireless Sensor Networks," in *Proc. 2nd Int. ACM Workshop on Wireless Sensor Networks and Applications*, San Diego, CA, Sep. 2003, pp.20-29.

[11] N. Katenka, E. Levina, and G. Michailidis, "Robust target localization from binary decisions in wireless sensor networks," *Technometrics*, vol. 50, no. 4, pp. 448-461, Nov. 2008.

[12] P. P. Joshi and T. C. Jannett, "Performance-guided reconfiguration of wireless sensor networks that use binary data for target localization," *Third International Conference on Information Technology, New Generations*, 2006, *ITNG 2006, pp*. 562-565.

A Proposal of a Nonlinear Observer Applied to the Heat Exchange Phenomena in a Double Pipe Heat Exchanger

J. F. Briones[1], M.A. Paz[1], J.L. Gallegos[1], J.I.O. Rodriguez[1], S.C. Garibo[2]
[1]Department of Electronic Engineering, Universidad Politecnica de Aguascalientes, Aguascalientes., Mexico
[2]Department of Electrical Engineering, Instituto Tecnologico de Aguascalientes, Aguscalientes, Mexico
Phone (449) 442-1400 Fax (449) 442-1419 E-mail: francisco.briones@upa.edu.mx

Abstract — **The aim of this paper is to show the heat exchange phenomena applied in a double pipe heat exchanger pilot plant scale, by considering a macroscopic dynamic model, that describes the evolution of the temperature of hot and cold fluids in the equipment. The proposed dynamic model is function of mean effective temperature difference and of the thermodynamic and transport properties, which are of strictly uncertain nature. These uncertainties can be isolated with the use of mathematical tools of differential geometry, based on Lie algebra and nonlinear dynamic observer design that uses only the information related with mass fluid flows and the mass contained in the equipment. The numerical simulations represent the heat exchanger physical behavior, and these results can be used to formulate robust control strategies physically feasible.**

Keywords — **Nonlinear observer, differential geometry, double pipe heat exchanger.**

I. INTRODUCTION

Because of their numerous applications in industrial processes, heat exchangers have been the subject of many studies including, among others: steady-state, transient, and frequency response analysis; open-loop qualitative behavior characterization; fault diagnosis/detection; numerical simulation; feedback control; parameter identification; and state reconstruction. Among these topics, the two lastly mentioned have played an important role in the formulation of solution that help cope with the operation conditions imposed to current industrial processes [5].

Particularly, the heat transfer phenomena involve different thermodynamic properties like: thermal dispersion coefficients by conduction and convention (e.g. film coefficients) and heat capacities associated with the fluids contained in the heat exchanger equipment, also the transport properties that affect the heat conduction by convection, specifically the boundary layer and viscosity of the fluids [8].

Other variables should be considered in the design of the equipment, such as: equipment geometry and material used in the equipment production; heat transfer regime (concurrent and countercurrent fluid flows); and the spatial configuration (e.g. double pipe or tube and shell heat exchanger).

These variables must be considered as part of heat exchange equipment design criteria [1,9], and also they mean uncertainties that can be estimated by assumptions at the moment when dynamic behavior of the equipment is formulated, together with other kind of variables, associated with the use of the equipment, like the fouling factor that decreases the efficiency of the equipment. The way this factor begins to form is uncertain, also the grade of affectation to the equipment through the time.

Regarding the dynamic model used to describe the states (temperatures) of the equipment, this model has a high nonlinear nature, because it involves different theoretical-experimental correlations used to calculate certain indicators such as: (i) the overall heat transfer coefficient; (ii) the mean effective temperature difference; (iii) and the heat capacity of the fluids used in the equipment.

Finally, it is necessary to mention that other proposals related with the design of observers applied to this kind of equipment have been advance (see [5,6,7]). In those cases, the proposed observers use the mean effective temperature difference and overall heat transfer coefficient, and in this work we will present a proposal of a nonlinear observer, that can describe, with a minimal margin of error, the dynamic behavior of the equipment temperatures, only by using the information reflected by the flows of fluid and by the mass contained inside the equipment; and without using the mean effective temperature difference nor the overall heat transfer; indeed, the designed nonlinear observer can estimate these parameters as uncertainties that can be reconstructed by the measuring of temperatures, since these parameters have a strong dependence on the temperature, as has been reported in [8].

II. DESIGN OF THE DYNAMIC MODEL FOR THE DESCRIPTION OF HEAT TRANSFER.

The dynamic model presented in this paper is formulated from a macroscopic heat transfer balance that involves the parameters described in the past section.

The proposed dynamic model is validated with the experimental information obtained from a double pipe heat exchanger pilot plant with concentric pipes [2]:

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_68, © Springer Science+Business Media B.V. 2010

$$\dot{x}_1 = \frac{2}{M_1}\left(F_1\left(x_{1e} - x_1\right) - \frac{A_i U_i}{Cp_1}\Delta x\right)$$
$$\dot{x}_2 = \frac{2}{M_2}\left(F_2\left(x_{2e} - x_2\right) + \frac{A_o U_o}{Cp_2}\Delta x\right) \quad (1)$$

Where $x \in \Re^2$ is a vector space that represents the temperatures in the heat exchanger; M is the mass contained in the equipment (i.e. the mass distributed in the equipment that can be calculated from the density relation -considered known- using the volume as the space that contains the heat or cold fluid in the inner tube or the annular space); F is the mass fluid flow; A is the heat transfer area; U is the overall heat transfer coefficient; Cp is the heat capacity of the fluid; and Δx is the mean effective temperature difference, that defines in the system (1) the heat transfer regime (concurrent and countercurrent fluid flows).

The subscript *1* represents the heat fluid; *2* represents the cold fluid; *e* represent the conditions of both fluids at the entrance of the equipment; *i* and *o* represents the reference point in the equipment for the calculus of the property (*i* is the inner tube surface and *o* is the outer surface of the inner pipe).

III. DOUBLE PIPE HEAT EXCHANGER PILOT PLANT.

The double pipe heat exchanger pilot plant contains an inner copper pipe and an outer steel pipe that has to be isolated, in order to avoid heat losses in to the environment.

The cooper pipe has an inner diameter of 11mm and an outer diameter of 12mm; the outer pipe has an inner diameter of 26mm.

The effective longitude for heat transfer in the equipment is 7.62m. The time necessary to reach the equilibrium between the temperatures is 5 to 30 seconds.

The equipment has a set of valves that allows controlling the flow of fluids and that predisposes the regime with which the heat transfer will be used in the heat exchanger (concurrent and countercurrent).

This set of valves determines the way in which the fluids are going to flow: hot water circulating inside the inner cooper pipe and cold water circulating in the annular space between the pipes.

The temperatures and the flows of both fluids are registered at the entrance and at the exit of the equipment to calculate: (i) thermodynamic and transport properties of the fluids; (ii) and the characteristic parameters related with the heat exchanger and the equipment (see Figure 1).

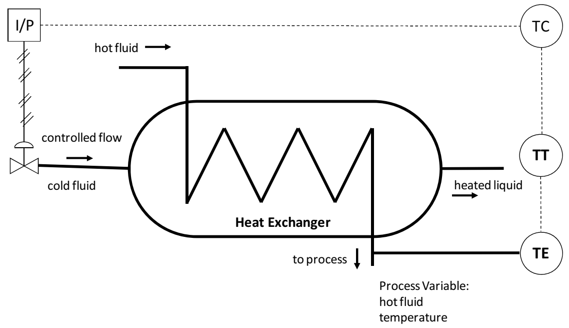


Fig. 1. Instrumentation diagram of the double pipe heat exchanger pilot plant

IV. PROPOSED NONLINEAR OBSERVER.

This design of a nonlinear observer uses the next proposal affine system:

$$\dot{x} = f(x) + g(x)u$$
$$y = h(x) \quad (2)$$

Where $y=h(x)=[x_1,x_2] \in \Re^2$ is the set of temperatures to be operated in the heat transfer, and u has been chosen as the mean effective temperature difference (Δx), because this term commonly is estimated as a logarithmic or arithmetic mean, but strictly this mean is uncertain inside the equipment and it is the main one among the others, like the heat capacity and the overall heat transfer. The vector maps *f(x)* and *g(x)* are defined by the system (1) as:

$$f(x) = \begin{vmatrix} f_1(x) \\ f_2(x) \end{vmatrix} = \begin{vmatrix} \dfrac{2}{M_1}(x_{1e} - x_1) \\ \dfrac{2}{M_2}(x_{2e} - x_2) \end{vmatrix}$$
$$g(x) = \begin{vmatrix} g_1(x) \\ g_2(x) \end{vmatrix} = \begin{vmatrix} \dfrac{-2A_iU_i}{M_1Cp_1} \\ \dfrac{2A_oU_o}{M_2Cp_2} \end{vmatrix} \quad (3)$$

If $y=h(x)=[x_1,x_2]$, it is easier to show that the mean difference temperature (independently from the heat transfer regime) is a function of y, and it allows for the uncertainty to be estimated, but before, it is necessary to establish the mathematical relationship between y and u.

Using Lie's algebra and the definition of the relative degree (ρ) [3], considering this degree the minimal whole number of derivates applied with Lie's algebra necessary to establish the mathematical relationship between y and the vector maps *f(x)* and *g(x)*, that is, $LgLf^k h(x)=0$ and $LgLf^{\rho-1}h(x)\neq 0$, where $k<(\rho-1)$, this combination must be validated for all $x \in \Re^2$.

The L operator is declared by [3]:

$$L_f h(x)=\sum_{i=1}^{n}\frac{\partial h(x)}{\partial x_i}f_i(x);\ L_f^k h(x)=L_f\left(L_f^{k-1}h(x)\right)(x)$$

$$L_f L_g h(x)=\sum_{j=1}^{n}\frac{\partial}{\partial x_j}\left(\sum_{i=1}^{n}\frac{\partial h(x)}{\partial x_i}g_i(x)\right)f_i(x)$$

Thus the relative degree of the system (1) is equal to 2, and due to its coincidence with the number of states in the system (1), then the system (1) has a *complete relative degree* and this guaranties that y can be used to describe the uncertain term and that there are no unmodeled dynamics.

The consideration of a complete relative degree for the system (1) guaranties the existence of a stable nonlinear observer, that will converge in its dynamic behavior with the system (1) when time has an infinite limit [4]:

$$\begin{aligned}\frac{d\hat{x}}{dt}&=f(\hat{x})+g(\hat{x})u+[Q(\hat{x})]^{-1}K[y-h(\hat{x})]\\ y&=h(x)\end{aligned}\tag{4}$$

Where the superscript ˆ denotes the observed states and $K=[k_1,k_2]^T\in\Re^2$ is the set of observer gains with magnitude and sign such as the states from system (4) tend to the states from the system (1) when time tends to infinity and the Jacobian matrix $Q(x)$ is defined by:

$$Q(x)=\begin{vmatrix}h(x)\\ L_f h(x)\\ \vdots\\ L_f^{n-1}(x)\end{vmatrix}$$

However, with the use of the vector maps $f(x)$ and $g(x)$ given by the system (3) onto the system (4), besides the Jacobian matrix above defined in terms of the Lie algebra and executing the rules associated to this kind of algebra, the nonlinear observer is:

$$\begin{aligned}\frac{d\hat{x}_1}{dt}&=\frac{2F_1}{M_1}(x_{1e}-\hat{x}_1)+\\ &\frac{M_1(2F_2k_1(x_1-\hat{x}_1)+M_2k_2(x_2-\hat{x}_2))}{2(F_2M_1-F_1(t)M_2)}\\ \frac{d\hat{x}_2}{dt}&=\frac{2F_2}{M_2}(x_{2e}-x_2)+\\ &\frac{M_2(2F_1k_1(x_1-\hat{x}_1)+M_1k_2(x_2-\hat{x}_2))}{2(F_1M_2-F_2M_1)}\end{aligned}\tag{5}$$

V. ANALYSIS OF THE DYNAMIC SYSTEM.

The analysis consists of the comparison of the numerical simulation between the dynamic behavior of the systems (1) and (5) in concurrent and countercurrent fluid flows.

In the system (5), the observation gains locate the characteristic polynomial poles in -1, where this characteristic polynomial is defined by the error system formed by the difference between the systems (5) and (1), and it allows for the dynamic behavior of the system (5) to tend to the dynamic behavior of system (1), when the time tends to infinity.

Apart from what has been mentioned before, it is necessary to consider only known nominal values of the parameters involved in the systems (1) and (5), and they are calculated at the heat exchange initial conditions in the equipment. These values are defined in Table 1.

TABLE I
NOMINAL VALUES USED IN THE NUMERCIAL SIMULATIONS IN CONCURRENT AND COUNTERCURRENT HEAT TRANSFER

Parameters	Heat exchanger regime	
	Concurrent	Countercurrent
Ui [=] kcal/m² ºC hr	115.033	100.244
Uo[=] kcal/ m² ºC hr	103.286	89.653
Ai[=]m²	0.2736	
Ao[=]m²	0.3040	
F_1[=]kg/hr	11.279	
M_1[=]kg	0.767	
M_2[=]kg	3.258	
Cp_1[=]kcal/kg ºC	0.999	
Cp_2[=]kcal/kg ºC	0.997	
x_1e[=]ºC	90	
x_2e[=]ºC	29	

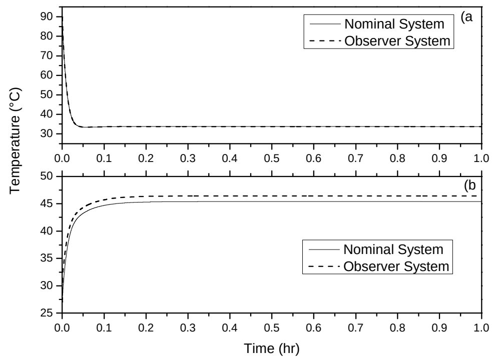


Fig. 2. Numerical simulation of heat exchange in countercurrent regime. a) Hot fluid temperature. b) Cold fluid temperature.

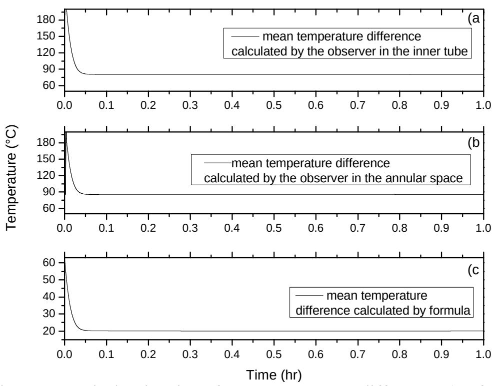


Fig. 3. Numerical estimation of mean temperature difference. a) reference by observation in the inner tube. b) reference by observation in the annular space. c) definition of logarithmic mean temperature difference.

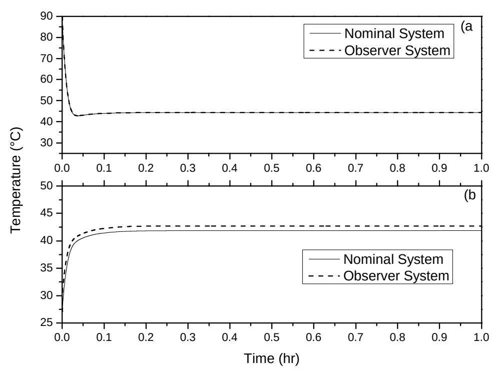


Fig. 4. Numerical simulation of heat exchange in concurrent regime. a) Hot fluid temperature. b) Cold fluid temperature.

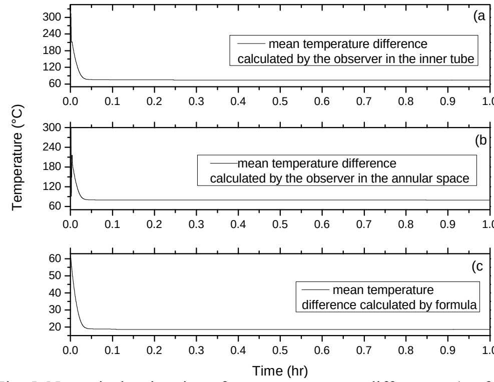


Fig. 5. Numerical estimation of mean temperature difference. a) reference by observation in the inner tube. b) reference by observation in the annular space. c) definition of logarithmic mean temperature difference.

Figure 2 and Figure 4 show that system (1) and (5) have an equal dynamic behavior in terms of temperature evolution over time, despite the fact that system (5) uses only mass and flow data from the equipment, meanwhile system (1) needs more information, particularly the mean effective temperature difference and the overall heat transfer coefficient.

However, the nonlinear observer (5) has a better description of hot temperature than cold temperature regardless of the heat exchange regime, because the difference between the real and the estimated hot temperature is practically zero, meanwhile the difference between the real and estimated cold temperature is 1°C.

Thus the nonlinear observer (5) is a promissory dynamic model in terms of describing the hot temperature in the double pipe heat exchanger using a minimum of information, and in physical applications, the hot temperature has more importance than the cold temperature, because the first one is commonly the variable that is intended to control in a process.

Figure 3 and Figure 5 show the dynamic evolution of the mean temperature difference, this way, the observer system describes the mean temperature difference in the equipment, regardless of heat exchanger regime, and this information is complementary to the measure of the maximum quantity of heat that the fluids can exchange.

This is an extra value, because this kind of nonlinear observer allows user to isolate the principal uncertainties in the heat transfer process (the mean effective temperature difference and the overall heat transfer coefficient), and eventually describes the properties of the equipment.

So, even if there is a logarithmic mean effective temperature difference (Figure 5c), this parameter only describes the capacity of temperature that can be transferred between fluids (approximately 20°C), but the estimation of this parameter obtained from the nonlinear observer, describes the internal temperature of the equipment (about 80°C), and this information is more useful if one of the process objectives is controlling the inner temperature in the equipment, to avoid the formation of fouling in the inner tube and the annular space and this measure is very important, because it is much too difficult to install a set of temperature sensors inside the equipment to obtain this information, which can be estimated by the proposed nonlinear observer.

VI. CONCLUSIONS

The proposed nonlinear observer is a promissory dynamic model to be used in the synthesis of advance control strategies, in which the only variables known are flows and masses in the equipment.

While there are other observers previously proposed ([5,6,7]), they use two critical parameters (the mean effective temperature difference and the overall heat transfer coefficient), that can be estimated from experimental correlations with some margin of error. The nonlinear observer hereby proposed does not use them and it even estimates them with a minimal margin of error than the experimental correlations estimate.

Even more, this nonlinear observer allows the user to estimate the temperature inside the equipment and uses this estimation for other applications besides control process (e.g. improve de the design of double pipe heat exchanger), and thus, it has with wide possibilities of being physically implemented.

REFERENCES

[1] Yunus A. Çengel, *Heat and Mass Transfer – A Practical Approach*. Singapore, McGraw-Hill, 2006, pp.619-626.

[2] Y.S.N. Malleswararao and M. Chidambaram, "Nonlinear controllers for a heat exchanger", *J. Process Contr.*Vol. 1, No.2, pp. 17–21, 1992.

[3] Alberto Isidori, *Nonlinear control systems: An introduction,* Germany, Springer, 1989, pp. 145-150

[4] H. Murata et. al., "Control of inverted pendulum-wheeled cart system using nonlinear observers", *SICE'95 Proceedings of the 34th SICE Annual Conference. International Session Papers*, Vol. 26-28, pp. 1133-1136, July 1995.

[5] A. Zavala-Rio et. al., "Bounded positive control for double-pipe heat exchangers", *Control Engineering Practice*, Vol. 17, No. 1, pp.136-145, January 2009.

[6] C.-M. Astorga-Zaragoza et. al., "Performance monitoring of heat exchangers via adaptive observers", *Measurement,* Vol. 40, No. 4, pp. 392-405, May 2007.

[7] C.-M. Astorga-Zaragoza et. al., "Observer-based monitoring of heat exchangers", *ISA Transactions*, Vol. 47, No. 1, pp. 15-24, January 2008.

[8] Prabhata K. Swamee et. al. "Optimum design of double pipe heat exchanger", *International Journal of Heat and Mass Transfer* , Vol. 51, pp. 2260–2266, 2008.

[9] T.W: Fraser Russell, et. al., *Mass and Heat Transfer – Analysis of mass contactors and heat exchangers,* United States of America, Cambridge University Press, 2008, pp. 90-102.

Vibration Signal Processing by Deconvolution Method

Popescu Ion Marian
University Constantin Brâncuşi Târgu-Jiu, Romania, marian@utgjiu.ro
Runceanu Adrian
University Constantin Brâncuşi Târgu-Jiu, Romania, adrian_r@utgjiu.ro

Abstract. **This paper presents a method of processing vibration signal provided by a piezoelectric sensor. The main purpose is to identify the actual vibration signal of a mechanical structure, and development of estimation methods with a certain precision for the real signal provided by sensor signal processing, using the deconvolution method. Basically, knowing the sensor output signal, knowing mathematical model of the sensor (with a certain precision of calculation) we are trying to find the input signal sensor, which is a real vibration signal.**

I. INTRODUCTION

A very important method for predictive detection is the measurement of defects in the mechanical vibrations that occur in these, relatively high accuracy.

By analyzing the acquired signals can develop efficient algorithms to detect a possible future failure even before they appear. To develop such a system was initially developed in an experimental model consisting of a DC motor which rotates a shaft with two bearings, one new and one with visible defects. Structure experimental model was developed is shown in Fig.1.

Vibrations are measured using piezoelectric sensors {4} whose output signal is applied to a signal conditioning circuit {3} to be then converted into digital signal using an acquisition system {2}.

Digital signal is saved and sent to a computer system {1} for processing and obtaining of information concerning the status of the system under observation.

In the stand the bearings under observation are connected on the same shaft that is rotated by a DC motor with speed up to 1500rpm, motor which is controlled by an electronic module.

In this application, was designed and developed a high-speed acquisition system to acquire a large number of samples, store them in the local memory, then to transmit the communication channel to a computer system at a much lower speed compared to the speed of acquisition.

This system allows off-line processing of the signal acquired by the computer system (1) but has the advantage to make a distributed acquisition system, that is placed near the mechanical system, then the remote can transmit digital data through a standard communication channel.

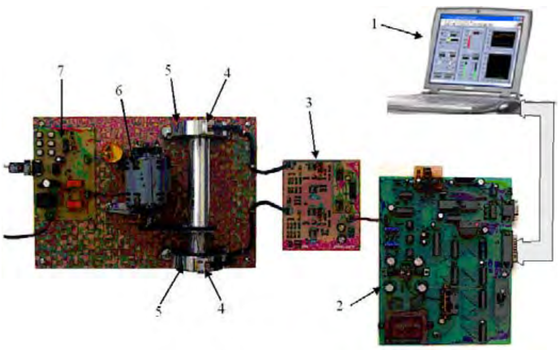


Fig.1.Experimental platform for vibration acquisitions

Acquisition system for signal processing from vibration sensors is designed around a core with microcontroller, that has an acquisition speed to around 400,000 Samples/second and the local store of a number of samples corresponding to an observation period of 2 ÷ 3 seconds.

The structure of the acquisition system is shown in Fig.2.

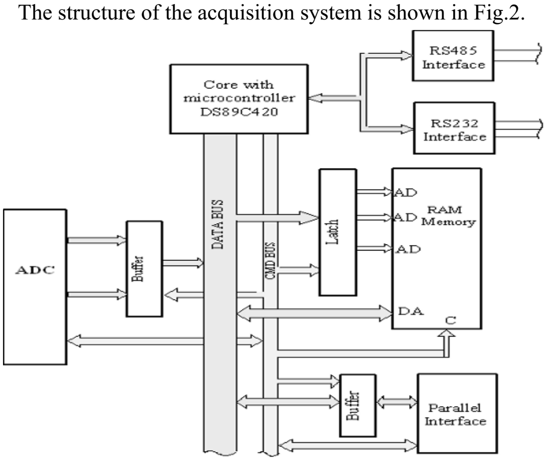


Fig.2.The structure of the acquisition system of vibration

The acquisition system includes a 12-bit analog/digital converter, a 2MØ static RAM, a standard serial port RS232

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_69, © Springer Science+Business Media B.V. 2010

that is the interface for programming and communications over short distances, a serial interface RS485 for long distances.

To qualify for the rate of acquisition was used for DS89C420 microcontroller core, which offers good performance and is compatible with standard microcontrollers family Intel8051.

This microcontroller has an improved structure and execute almost any instruction from the instruction set of microcontroller Intel8051 up to twelve times faster at same clock frequency of the system.

II. MODELING THE VIBRATION TRANSDUCER

Overall construction vibration transducer included 2 elements: a mechanical element, which follows the motion of the vibrating structure and a sensor, which converts relative motion into an electrical signal output.

Vibration transducers type seismic devices consist of a mass-spring-damper system, connected by a fixed support structure of the vibration and measuring the relative movement of mass to support.

Basics elements of a seismic sensor are shown in Fig.3.

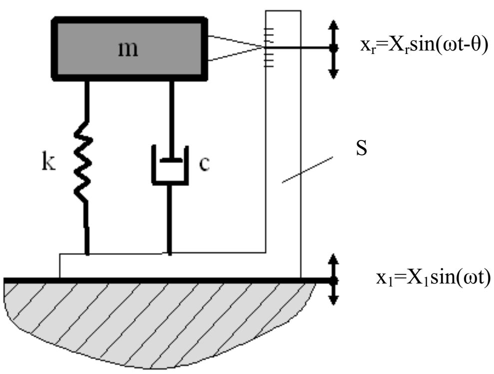


Fig.3. The principle of a vibration transducer

Support device S is connected rigidly to the structure of mechanical vibration. Mass m is attached to transducer support through spring k and damper c. We will consider that the support device executes a harmonic motion:

$x_1 = X_1 \sin(\omega t)$

Relative movement of the mass to the support is the response of the transducer:

$x_r = X_r \sin(\omega t - \theta)$

Mass displacement relative to a fixed reference point in space is: $x = x_1 + x_r$ and the corresponding acceleration is $\ddot{x} = \ddot{x}_1 + \ddot{x}_r$. Applied forces acting on the mass of spring and damper, proportional with x_r-movement and respectively $\dot{x}_r$-speed, which are elastic force that viscous damping force.

The equation of motion of mass m is therefore:

$$m\left(\ddot{x}_1 + \ddot{x}_r\right) + c\dot{x}_r + kx_r = 0 \tag{1}$$

or

$$m\ddot{x}_r + c\dot{x}_r + kx_r = -m\ddot{x}_1 \tag{2}$$

Using these notations and equation solving earlier[4], obtain the relationship:

$$\frac{X_r}{X_1} = \frac{\left(\frac{\omega}{p}\right)^2}{\sqrt{\left[1 - \left(\frac{\omega}{p}\right)^2\right]^2 + \left(2\xi\frac{\omega}{p}\right)^2}} \tag{3}$$

$$\theta = arctg\frac{2\xi\frac{\omega}{p}}{1 - \left(\frac{\omega}{p}\right)^2} \tag{4}$$

where: $p = \sqrt{\frac{k}{m}}$, $\xi = \frac{c}{c_{cr}}$, $c_{cr} = 2\sqrt{km}$

Relationship (3) is simulated with the software package Matlab 7.1 in Fig.4. In this figure both curves tend to value $X_r/X_1=1$ to frequencies ω>>p. In part III of graph $X_r \cong X_1$, so the relative motion between the seismic mass and support is the same movement support.

In Fig.5. will result in this domain, for a weak damping, phase angle is $\theta \cong 180^0$, so the mass *m* and support vibrating in opposition . Mass *m* is a fixed point to which the support device has a harmonic motion equal to that structure. If the transducer is moving, the signal is proportional to the displacement of the structure studied, and is called vibrometer seismic transducer.

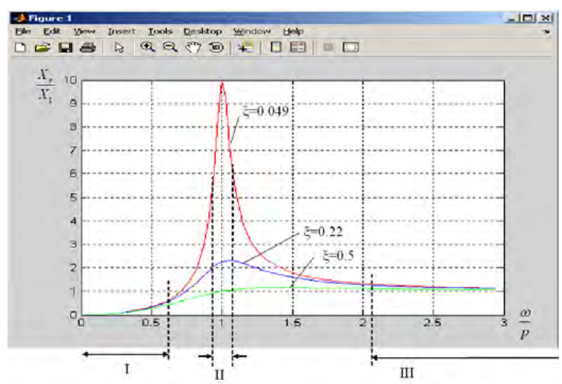


Fig.4. Amplitude-frequency characteristic

Because the natural frequency of the "mass-elastic suspension" must be much smaller than the measured vibration frequency, displacement sensors have a "soft" suspension. This makes the relative movement of the seismic mass is greater at lower frequencies. Consequently, displacement sensors have relatively large dimensions and weight, to be taken into account when measuring small mechanical structures, whose response can be so greatly influenced.

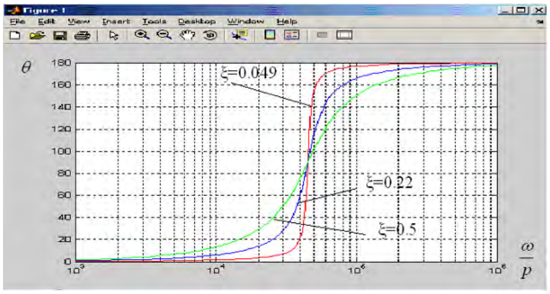


Fig.5.Phase-frequency characteristic

To frequencies $\omega << p$ (interval I of Fig.4.) relationship (3) becomes: $\frac{X_r}{X_1} \cong \left(\frac{\omega}{p}\right)^2$ so that we have $X_r = \frac{1}{p^2}\left(\omega^2 X_1\right)$ where the expression in brackets is the amplitude of the absolute acceleration of the box.

Movement of the seismic mass is in phase with that of the support. If one neglects damping, spring deflection (i.e. the sensor response) is proportional to the force of inertia of the mass and proportional to the acceleration of mass. Results in this operating mode, output signal of a seismic displacement transducer is proportional with acceleration structure studied (p^2 is a constant device). We can say the device works as a seismic accelerometer. If the transducer is the speed at the same operating conditions, device measures the acceleration of the order 2 ("jerk") of the structure. Unlike displacement sensor, seismic accelerometer has a relatively strong spring, so a relatively high natural frequency and a total weight much smaller than the vibrometers. To frequencies $\omega \cong p$, seismic device working in resonance zone II, the indications of this become very large and the phenomenon is used only to measure the frequency.

Generaly, vibratory motion of the structure studied is not harmonic, as assumed above. When measuring a not harmonic movement, the sensor response to different spectral components is different.

As I shown, the presence of damping, amplification X_r/X_1 and phase angle θ, varies with the report $\frac{\omega}{p}$, so that the device changes the report of amplitude and phase between different harmonics, making the resultant movement is distorted by the real movement.

As a result, seismic devices with high damping give rise to distortions when measuring not harmonic vibration.

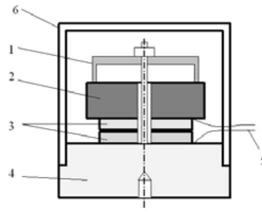
The most commonly used accelerometers used as piezoelectric transducers.

Principle scheme of a seismic sensor, with piezoelectric crystal used as an elastic element, is shown in fig.6.

Due to high rigidity of crystals, natural frequency of the sensor is high, so it can be used only as an accelerometer. When attached to a body that vibrate with the acceleration *a*, seismic mass *m* acting with a force *F* = *ma* on the piezoelectric disc.

Due to mechanical solicitation, the internal electrical echilibrium of the piezoelectric materials is disturbed, and to restore him, on the 2 sides opposite is deposited electric charges.

It produces a variable voltage (dependent electrical capacity of the system), proportional to the force, so the acceleration of mass, which can be measured.



1- elastic tensioning element
2- seismic mass
3- piezoelectric element
4- the sensor base
5- Electrical connections
6- case

Fig.6.Piezoelectric vibration transducer

For realizing the conditions for the experimental model was chosen MAQ36 accelerometer, because has its own natural frequency of such large and useful field measurement of the vibration frequency is higher.

MAQ36 produced by SENSOTEC company, is a miniature accelerometer, the output signal is in charge, designed for use in industrial applications and automation, including experimental models that require little mounting space and high natural frequency. MAQ36 is a self generating piezoelectric transducer, which has no internal electronics and requires no external power for operation.

The accelerometer should be connected to an charge amplifier, placed as near as possible to the transducer in the measuring chain. MAQ36 is an compression accelerometer, which due to small size ensures a high own natural frequency.

Construction of its dimensions are shown in Fig.7.

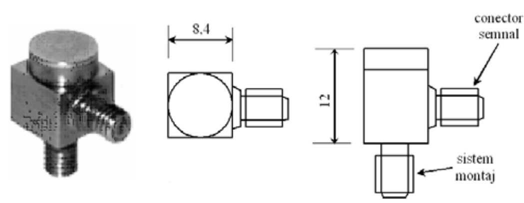


Fig.7.Accelerometer MAQ36

The main features of the accelerometer (to the catalog) are:

- Sensitivity 5pC/g ±10%
- The frequency of work 1Hz until 30KHz
- Natural-frequency assembly 45KHz
- Dynamic range ±2000g
- Sensitivity to temperature 0.08/°C
- Transverse sensitivity < 5%
- Amplitude linearity better than 1%
- The temperature range -90°C +900°C
- Capacity 300pF
- Weight 10 grams
- Force Mounting 9N

III. PROCESSING THE ACQUIRED VIBRATION SIGNAL

Using data acquisition system [1] described above and generate vibration platform were obtained the following signals acquired for a bearing "good" and a bearing “spent” as in Fig.8.

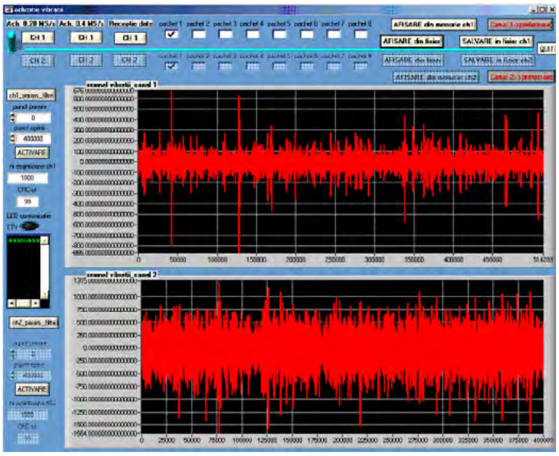


Fig.8.Vibration signals (acceleration) acquired

The acquisition was done for 1 second time and have obtained 400,000 samples. Signals acquired for the 2 bearings that made a zoom for 1000 samples is presented in Fig.9.

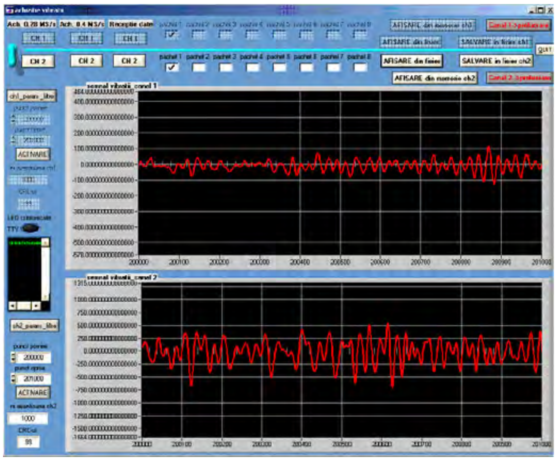


Fig.9. Vibration signals for the 2 bearings consist 1000 samples

Processing was realized using the development environment LabWindows/CVI 8.1 and processing was done for the acquired signal from the charge amplifier, that is connected accelerometer. The signal is a signal of acceleration (vibration) of the mechanical system.

We made for these 2 signals the power spectrum (using Fourier transform) and obtain the processing of Fig.10.

As you can see, most of the vibration signal energy is at frequencies above 1 kHz or even 10kHz.

The power spectrum for the “good” bearing is 20 times lower than the power spectrum for the “spent” bearing. As the motor that rotates the mechanism of the 2 bearings that are fitted vibration sensors have 1500rpm, i.e. around 25Hz, it can be concluded that in this signal is included vibration and effective response vibration sensor which has a important weighting.

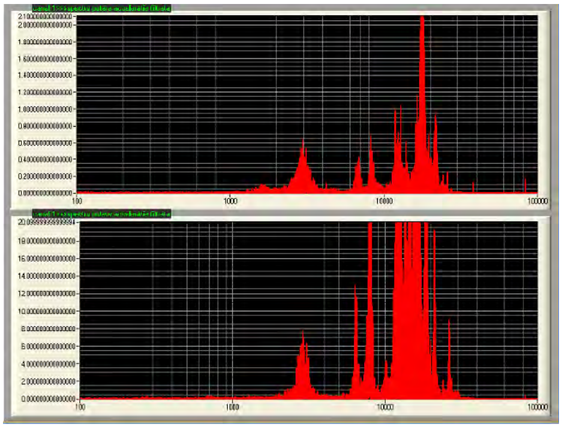


Fig.10. Power spectrum of acquired signals

Basically, the response of the transducer $x_r(t)$ is obtained as the convolution product (in case of a linear and time invariant system) between signal $x_1(t)$ and weighting function h(t) of the vibration transducer, where $x_1(t)$ is the real vibration signal.

$$x_r(t) = x_1(t) * h(t) \tag{5}$$

or:
$$x_r(t) = \int_{-\infty}^{\infty} x_1(\tau) h(t-\tau) d\tau \tag{6}$$

for causal $x_1(t)$ and h(t), the convolution integral limits can customize as follows:

$$x_r(t) = \int_{0}^{t} x_1(\tau) h(t-\tau) d\tau \tag{7}$$

Returning to equation (2), we have:

$$m\ddot{x}_r(t) + c\dot{x}_r(t) + kx_r(t) = -m\ddot{x}_1(t) \tag{8}$$

Starting from the premise that mechanical vibration of the structure is small displacement around a position of equilibrium which is also a stationary point for the system, approximating the vibration transducer as a linear system is valid.

So, we can apply the Laplace transform and we have zero initial conditions:

$$ms^2X_r(s) + csX_r(s) + kX_r(s) = -ms^2X_1(s) \tag{9}$$

Mathematical model (transfer function) of the vibration transducer is:

$$H(s)=\frac{X_r(s)}{X_1(s)}=\frac{-ms^2}{ms^2+cs+k} \tag{10}$$

Using the Tustin substitution:

$$H(z)=H(s)\Big|_{s=\frac{2}{T}\frac{z-1}{z+1}} \tag{11}$$

$$\Rightarrow H(z)=\frac{-m\dfrac{4}{T^2}\dfrac{(z-1)^2}{(z+1)^2}}{m\dfrac{4}{T^2}\dfrac{(z-1)^2}{(z+1)^2}+c\dfrac{2}{T}\dfrac{(z-1)}{(z+1)}+k} \tag{12}$$

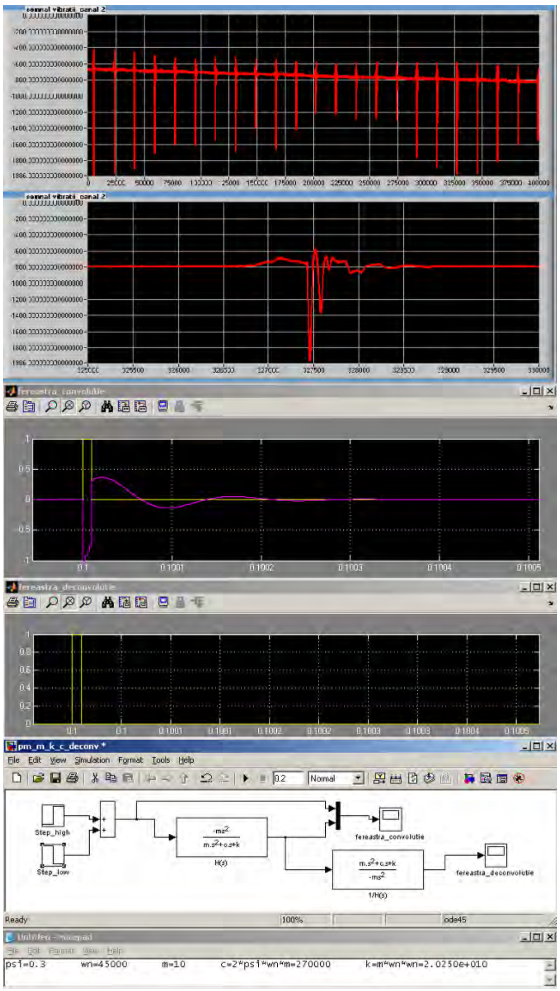


Fig.11. Rectangular impulse response for the real system and simulated system

After processing relationship (12), we have:

$$H(z)=\frac{X_r(z)}{X_1(z)}= \tag{13}$$

$$=\frac{-4m+8mz^{-1}-4mz^{-2}}{\left(4m+2Tc+kT^2\right)+\left(-8m+2kT^2\right)z^{-1}+\left(4m-2Tc+kT^2\right)z^{-2}}$$

Further we note this:

b_1=-4m; b_2=8m; b_3=-4m;

$a_1=4m+2Tc+kT^2$; $a_2=-8m+2kT^2$; $a_3=4m-2Tc+kT^2$

We obtain discrete transfer function:

$$H(z)=\frac{b_1+b_2z^{-1}+b_3z^{-2}}{a_1+a_2z^{-1}+a_3z^{-2}}=\frac{X_r(z)}{X_1(z)} \tag{14}$$

Equation (14) can be implementated by considering the delay operator z^{-1}:

$$a_1x_{rk}+a_2x_{rk-1}+a_3x_{rk-2}=b_1x_{1k}+b_2x_{1k-1}+b_3x_{1k-2} \tag{15}$$

where:

x_{rk}- is the signal of the output vibration transducer (transducer response)

x_{1k}- is the signal of the input vibration transducer (real vibration of bearing)

Real vibration signal for the mechanical system after this processing is:

$$x_{1k}=-\frac{b_2}{b_1}x_{1k-1}-\frac{b_3}{b_1}x_{1k\_2}+\frac{a_1}{b_1}x_{rk}+\frac{a_2}{b_1}x_{rk-1}+\frac{a_3}{b_1}x_{rk-2} \tag{16}$$

Because the transducer transfer function from relation (10) is a proper function, can be implemented numerically as relationship (16), we obtain the signal of input (real vibration of mechanical structure) of the transducer, knowing output signal of transducer and the mathematical model of the transducer. In this way, is made a digital deconvolution, which is valid given that the transfer function is proper.

In fig.11. first 2 graphs are the response of the transducer to impulse mechanically generated and the next 2 graphs represent the response of the mathematical model simulated.

Power spectrum obtained for the new signal x_1 is presented in fig.12. for the 2 bearings. As you can see, most of the signal energy is in the low frequencies.

Although the 2 vibration signals are acquired on different channels from acquisition board and from different bearings, can be observed a correlation of spectral components (highlighted by green dotted lines) in relation to frequency and amplitude compared with the notes that the bearing used (2nd graph) has some component of 200 times the bearing something better (first graph).

This method, with digital deconvolution, introduces large errors primarily due to initial conditions is not identical with real conditions.

In the graphs presented, after the processing by the relation (16) vibration signal, has eliminated a total of 100,000 samples from 400,000 before taking a Fourier transform. By this was removed from the calculated signal variation due to the initial conditions.

Another error that affects processing is not known with precision mathematical model of the vibration transducer and in particular the coefficients c and k which will have corresponding ω_n şi ξ . In this case it was considered m=0.002; ξ=0.1; ω_n =45000;

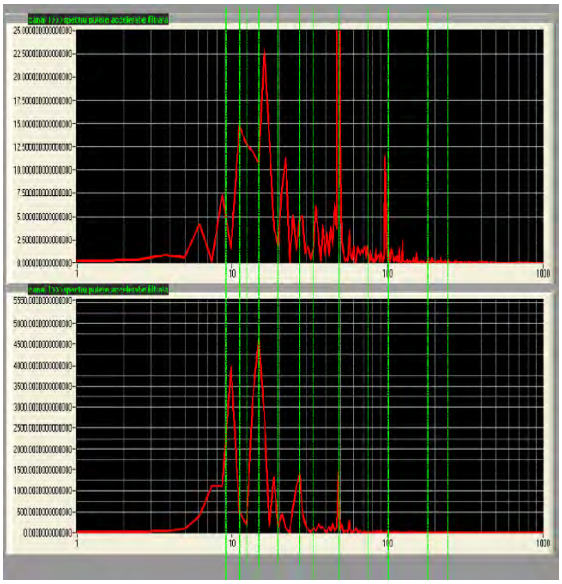


Fig.12. Power spectrum obtained after deconvolution

IV. Conclusions

In an automation system, digital signal processing of vibration is a very difficult problem which must take into account the following aspects:

> Acquisition system must meet the powerful features of real time, because the vibration signal is a very dynamic signal.

> Another problem is strongly nonlinear nature of a vibration signal, which complicates its processing. Most mechanical systems that generate vibrations have imperfections in their structures, to which adds that the vibration in a mechanical structure have more degrees of freedom and transmission through the structure.

> Vibration transducers have limited responses and introduce distortions of amplitude and phase signal provided.

> Any acquisition system (including transducer) and signal processing of vibration require laborious calibration platforms specially designed for a specific type of application.

> Structures analysis and on-line processing of vibration signals are very difficult to achieve because in a mechanical structure, the point where vibration measurements is made, is far from processing system, and networks cable introduce additional distortions

As a conclusion, the most applications for acquisition and processing of vibration in purpose of predictive maintenance of a mechanical structures, processing the signal for a mechanical structure "good", and this process taking as reference for comparison to other similar structures in operation.

References

[1] Grofu F., Popescu L. and Cercel C. (2008) *"Acquisition System for Monitoring Vibrations",* International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 2007) Conference**,** Conference Proceedings book, December 3-12, 2007, University of Brigeport, USA, publish in Novel Algorithms and Techniques In Telecommunications, Automation and Industrial Electronics, ISBN 978-1-4020-8736-3 (Print) 978-1-4020-8737-0 (Online), pg. 89-92, DOI 10.1007/978-1-4020-8737-0_16.

[2] Shin K., Hammond J., *"Fundamentals of Signal Processing for Sound and Vibration Engeneers",* John Wiley&Soons Ltd., 2008, ISBN 978-0-470-51188-6

[3] R.Keyth Mobley,"*An Introduction to Predictive Maintenance*", 2nd Edition, 2002, Elsevier Ltd., ISBN: 978-0-7506-7531-4

[4] Mateescu A.,Dumitriu N., Stanciu L. "*Semnale si sisteme-Aplicatii in filtrarea semnalelor*", Ed.Teora, 2001, ISBN:973-20-0666-8

[5] Buzdugan G., Fetcu L., Rades M. *"Vibratiile sistemelor mecanice"*, Ed. Academiei Romane Bucuresti, 1975.

[6] Moheimani**,** S.O. Reza, Fleming**,** Andrew J., *"Piezoelectric Transducers for Vibration Control and Damping"*, Series: *Advances in Industrial Control*, 2006, Springer-Verlag, ISBN: 978-1-84628-331-4

[7] Harris C.M.; Piersol, A.G. "*Harris' Shock and Vibration Handbook*", (5th Edition), 2002, McGraw-Hill, ISBN 0-07-137081-1.

[8] Douglas Thorby "*Structural Dynamics and Vibration in Practice*", 2008, Elsevier Ltd., ISBN: 978-0-7506-8002-8.

[9] Dan Necsulescu, "*Advanced mechatronics: monitoring and control of spatially distributed systems*", 2008, World Scientific, ISBN 9812771816, 9789812771810

[10] www.leonovabyspm.com, Leonova Infinity-Instruction Manual, 2008

[11] http://www.ni.com/soundandvibration/, *Sound and vibration systems,* 2009.

[12] Honeywell Sensotec, *Accelerometers Frequently Asked Questions*, www.sensotec.com, 2003

[13] Maxim-IC, *App Note 604: Fast Memory Transfers with the Ultra High-Speed Microcontroller - DS89C420,*

[14] I.J.Taylor. *"Identification of Bearing Defects by Spectral Analysis".* Journal of Mechanical Design, Transaction of ASME, vol. 120, 1980, pp 199-204.

[15] *Vibration Measurement and Analysis Brüel & Kjaer* Handbook BA 7676-12, 1998

An IT Change Management Tool to Optimize Scheduling and Planning

Denílson Cursino Oliveira, Raimir Holanda Filho
University of Fortaleza- UNIFOR
denoliveira@edu.unifor.br, raimir@unifor.br

Abstract – The IT governance has become strategic to business objectives. The main reason is the increase of importance that IT has in business decision. To achieve this importance, the IT management must ensure a Just-in-time feedback, electing the change management as main point of IT to help the business goal through the use of tools with optimized procedures to guarantee a lower financial loss during the implementation of changes. This work presents a tool to minimize the impact of the changes on the business goals, respecting the dependencies between the changes and featuring innovations as dynamic change window determination. The tool presented in the work builds an optimized determination of change windows, treatment of dependencies and allocation of changes, helping the business to minimize the impact.

I. INTRODUCTION AND MOTIVATION

The actual IT governance aligns the real goals of the business with an automated response information technology to achieve strategic goals with more efficiency and less time. To formalize the IT governance globally, some set of procedures and good practice have been developed. Control Objectives for Information and related Technology (COBIT) [1] is one of them that promotes a management and control of information with use of measures, performance indicators and maturity models. Other example is the IT information library (ITIL) [2], which consists of a set of best practices that determine how the IT infrastructure of a company is organized. The library ITIL is an IT service management framework that promotes an efficient management of IT changes in order to minimize any impact upon IT services.

Some automated IT tools were developed, using the good practice and procedures of ITIL and COBIT, aligning IT and business strategy, measuring the IT performance and building IT feedbacks to strategic decisions. We propose a change management tool to select the needed changes, determine in an automated form the best time intervals to become change window and build an optimized schedule of changes.

The first results of this study were published in [3], where we presented: the division of the changes in family of changes, where each one represents an IT service; the concepts of direct and indirect dependencies between the changes; the treatment realized over the dependencies and the rules to build an optimized schedule of changes based in one change window. The main feature of this model is the possibility of building a parallel implementation of changes using a simpler heuristic with same complex set of changes if compared with the state-of-art.

In [4] was added some resources to change management tool: more dependency treatment as the principle cross dependency; the allocation of changes in different change windows; the automated determination of time implementation of each family of changes, enabling the determination of the financial loss.

In [5] is presented an optimization of IT changes, showing mathematical concepts to determine the financial loss, cost of changes implementation over a set of changes attached to a change window. One change management conceptual model of a decision support tool is presented. The implementation of changes is performed sequentially. On the other hand, our model does this in parallel.

The CHAMPS [6] focuses on optimization of a scheduling of changes to satisfy external RFC´s. CHAMPS develops a complex dependence hierarchy that is used to build a final scheduling, where it reflects a least financial impact during the changes. Our tool has a simpler but powerful heuristic for the construction of the dependences between changes and their treatments. Also it builds an optimized scheduling based on financial or timeline variable.

The results obtained in previous work [3][4] showed that formalize some treatments over the change dependence and a simpler heuristic to build an optimized allocation of changes, now we present more features to the tool and a conceptual model.

II. THE CHANGE MANAGEMENT TOOL

As commented in previous work, as important as doing a planned change management is the selection of changes that should be implemented. We present a model divided into three main steps: change filtering, refinement of changes and scheduling.

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_70, © Springer Science+Business Media B.V. 2010

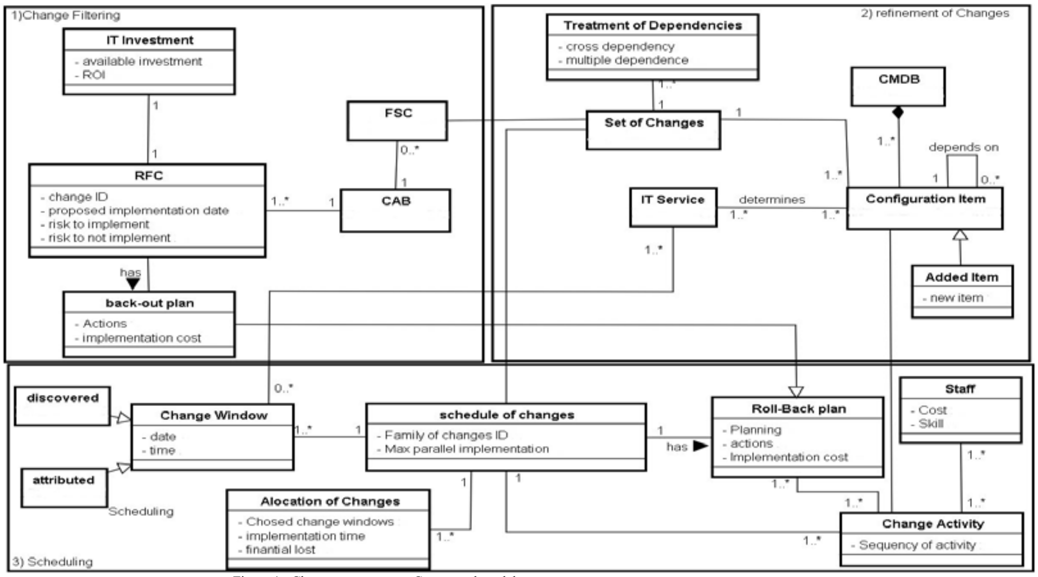


Figure 1: Change management Conceptual model

A. *Change Filtering*

The first step, we perform a filtering of all possible changes, represented by RFC documents. This document is a formal proposal to implement a change, referencing details as implementation cost, implementation and not implementation risks, proposed date to implement, etc.

IT investments adds details to RFC documents and the change advisor board (CAB) [2] analyze inherent to the proposed changes, considering the financial reserve to IT investments, return of investments (ROI) and other financial analysis. We want to propose a more detailed analysis with respect to ROI linked to this tool.

The back-out plan is formed by all actions to return the proposed change to your previous state in case of failure.

The CAB is formed by stakeholders from each area affected by the proposed changes. This group is responsible for receiving, analyzing, allowing or denying the changes. The set of changes allowed is described in forward schedule of changes (RFC).

In [4], we described that each change is represented by $C_{(i, j)}$, where i={1, 2, 3,…, n} and j={1, 2, 3, …, m}. The values of i and j will determine the position of $C_{(i, j)}$ in the scheduling of changes. The set of changes is represented by SC={$C_{(1, 1)},\ldots,C_{(n, m)}$}. The SC is formed by one or more families of changes. A family of changes is formed by changes $C_{(i, j)}$ with the same i and different value of j. In the Schedule, each family represents a column. The indirect dependence (ID) represents the dependence among changes with different i. We have a direct dependence (DD) when the implementation of $C_{(i,j)}$ depends on of $C_{(i,j-1)}$.

B. *Refinement of Changes*

In this step, all the changes have already been released. Our model requires treatments of these changes. These changes are linked to one or more configuration item (CI), which are part of a CMDB. The CMDB has a mapping of all CI, their dependencies and characteristics. Each change could be linked to new CIs (configuration item), which are out of CMDB. If this occurs, is necessary to map it to determine the dependencies between this new CI and the other that are part of CMDB.

After the mapping of dependencies between the changes, formation of the family of changes, our model applies treatments over the dependencies. The model provides two cases: cross dependency and multiple indirect dependencies from one family of changes to another [3] [4].

The final object of phase is the set of changes with all CIs and the mapping and treatments of your dependencies.

C. *Scheduling*

The schedule of changes centers the action implemented by the model to build an optimized allocation of the changes, considering a planning to get a minor implementation time and/or less financial loss.

The roll-back plan contains the actions to be implemented and gets the previous state of one change in case of failure during your implementation. The roll-back plan is different, bigger and more complex than back-out plan, because it includes adding, modifying or removing some of them. It modifies the action, takes into consideration two main points to do: The treatments of dependencies in the previous step and new mapping of dependencies by the addition of new CIs. Our studies showed that some modifications in the change actions are necessary to form the roll-back plan, although it inherits some of them from back-out plan.

The change windows are formed by two types: attributed and discovered change windows. Following the actual state-of-art, the attributed change windows are found negotiating with a client the time of this change window that have less impact for business objectives. If one or more changes affect more than one client, should be determined the change window considering these clients. In this case, which the less financial impact is more important, the chosen interval to become a change window is in periods of less intensity of work.

The discovered change windows are proposed by our model, in which the determination of interval chosen to become a change window is carried through dynamic form. If the stronger point to determine the interval is the impact in financial loss, we propose some steps to do it.

In our model, each family of change represents an IT service, in which we can determine the amount of financial profits they generate by a time interval. The interval used to generate the amount of financial profits is determined by the planned implementation time of the family of change linked to IT service. The choice of an interval can consider the reading of just one day or measuring of several. The chosen interval will be the one that generate minor profits (less financial impact). To illustrate the determination of a discovered change windows, we will use the implementation of the set of changes below.

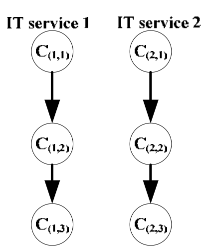


Figure 2: Set of changes

The figure 2 present two families of changes where each change has accurately one hour to be implemented, totalizing three hours to finish each family and consequentially the IT service linked to each one comes up. The tool would evaluate the profits of each IT service using a variable previously established, as showed in table 1.

Table 1. Analysis of financial profits by interval time.

	Intervals			
IT Service	0h to 3h	3h to 6h	...	21h to 24h
1	$ 1.200	$ 1.400	...	$ 2.100
2	$ 1.400	$ 1.300	...	$ 1.700

Analyzing the table 1, we have the change windows shown in table 2.

Table 2. Change windows.

IT Service	Intervals (hour/Day)	
	0h to 3h/ -	3h to 6h/ -
1	x	
2		x

In example of table 2 the implementation time of each family occurs in a sequential form, but it could be done in a parallel model. It is possible because they don't have indirect dependency between them.

The action of getting amount of profits by interval of time is one customizable point to tool. We could have an example where we need a customization to get the value in a just-in-time form.

The allocation of changes details all possible information about the chosen schedule. Here we have the final point of the model, in which the changes had been allocated to change windows and their implementation sequences determined.

III. A REAL-WORLD INSPIRED PROBLEM

Our model was implemented in many cases to validate it's your performance. The real case of figure 3 was implemented in an e-commerce company. The company has offices and needs to perform the following changes: Change the core-routers and switches for models belonging to the same manufacturer and capable of handling 802.1x packages; make changes in servers to fit the new authentication service. The company wants to deploy authentication of all computers and users to use network resources through 802.1X protocol.

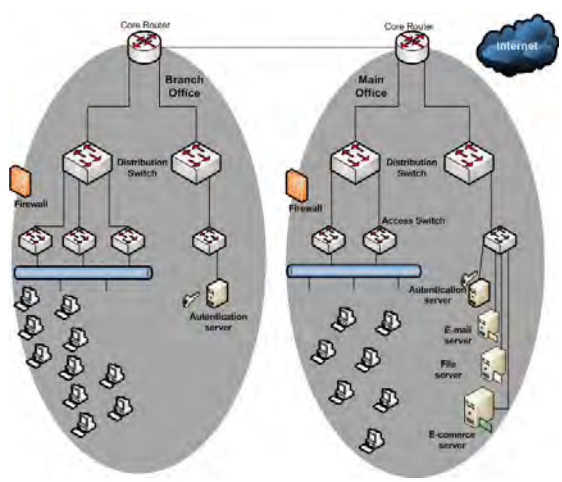


Figure 3: scene to implement the changes.

The branch office depends on the main one, because of this point, the exchange of core routers could be performed in parallel form. When the change of core-router of main office begins, the IT services of branch office break at the same time. The changes of core-switch were divided into three phases: change each unit; configure the operational system; perform security settings.

The change of switches could be performed just after the first and second phases of the core-router changes.

The changes of servers can start after the second phase of core-routers and distribution-switches. This is necessary because the change of servers to work with the authentication service depends on the changes of the routers and switches.

The model brings together the changes linked to an IT service to a column: the family of changes. In this case, to illustrate it, we consider the changes of each office as two different IT services. The table 3 shows the identification of IT service as family of changes.

Table 3: Identifying the IT services.

		Family of Changes (columns)					
		1	2	3	4	5	6
IT service	Server (Branch office)	X					
	Distribution switch (Branch office)		X				
	Core-Router (Branch office)			X			
	Core-Router (main office)				X		
	Distribution switch (Main office)					X	
	Server (Main office)						X

The identification of IT services to families does not follow a rule, therefore it is possible to identify them with any value of i. The representation of schedule of changes, implemented by our model, is in figure 4 below.

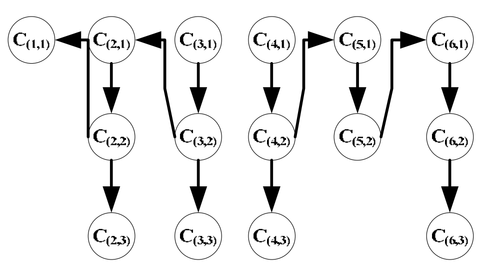


Figure 4: Schedule of changes.

The model takes the schedule of changes and applies the treatment of dependencies, verifying the existence of cross dependencies and multiple ID. After that, the rules of allocation of changes on array [4] are applied. In figure 5 shows the last state of schedule of changes.

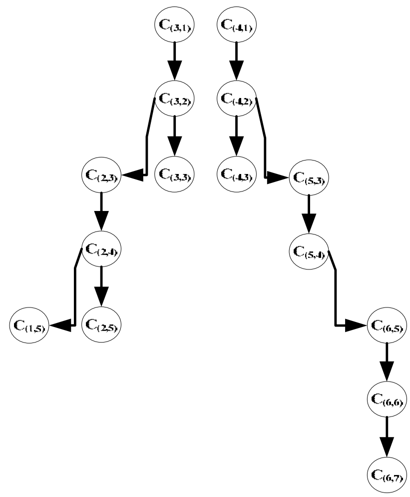


Figure 5: Schedule of changes.

It is important to note the fact that of family of change 6 (servers of main office) can only begin when the distribution switches are finished because the changes of servers take effect only if the other succeeds.

In this case, the change window was attributed and only one was necessary to implement all changes, with the interval was of 12h and was not necessary to use the resource of discovered change window of our model. The model has the function that calculates the total implementation time in just-in-time form [4]. Considering the implementation time necessary to each change of one hour, the total implementation time of each family is shown in table 4.

Table 4: Total implementation time of each family of changes.

Family of changes	Implementation Time
1	1
2	3
3	3
4	3
5	2
6	3

Just the changes of servers located at the main office generate financial loss. With an implementation time of three hours and rate of \$ 1.200/hour, the total financial loss was \$ 3.600. The other changes have an administrative functions without financial loss rate.

The model allocates the change in an array to facilitate the timeline vision of the implementation of changes as shown in table 5, following the last state of schedule of change of figure 5.

Table 5: Changes allocated in array.

		i					
		1	2	3	4	5	6
j	1			$C_{(3,1)}$	$C_{(4,1)}$		
	2			$C_{(3,2)}$	$C_{(4,2)}$		
	3		$C_{(2,3)}$	$C_{(3,3)}$	$C_{(4,3)}$	$C_{(5,3)}$	
	4		$C_{(2,4)}$			$C_{(5,4)}$	
	5	$C_{(1,5)}$	$C_{(2,5)}$				$C_{(6,1)}$
	6						$C_{(6,2)}$
	7						$C_{(6,3)}$

The implementation of changes follows two basic rules: all changes with the same value of i must to be implemented in a sequential mode and all changes with the same value of j could be implemented in a parallel mode [4]. The figure 6 shows the implementation of changes into an attributed change window.

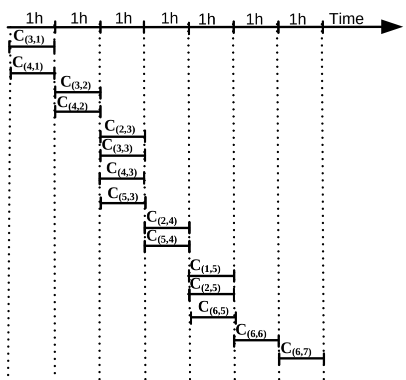


Figure 6: Schedule of changes.

IV. CONCLUSION AND FUTURE WORK

The results obtained through application of the model in many cases show the ability to offer an automated response to change management. The main feature is a simpler heuristic, if compared to others, to build a best schedule of change and planning your implementation for a smaller financial loss.

This paper presents a conceptual model of the tool with new functions as: discovering change windows, which can discover the time interval in an automated form; separating the concept of back-out plan and creating the roll-back plan, which must be more complex and complete.

We plan to develop an automatic reading of the schedule of change to build a roll-back plan. We will develop a detailed construction of roll-back plan to ensure the return of each configuration item to its previous state through an automated form.

REFERENCES

[1] IT Governance Institute, "COBIT 4.1 Edition, 2006, www.isaca.org/cobit.

[2] IT Infrastructure Library (ITIL 3.0), www.itil-fficialsite.com.

[3] Denilson Cursino de Oliveira, Raimir Holanda Filho, "A Highly parallel scheduling model for IT change management", in CISSE 2008.

[4] Oliveira, D.C.; Filho, R.H., "A time and financial loss estimation using a highly parallel scheduling model for IT change management", in Integrated Network Management-Workshops, 2009. IM apos; 09. IFIP/IEEE International Symposium on Volume, Issue, 1-5 June 2009 Page(s):1 – 9.

[5] Rebouças, R., Sauvé J., Moura A., Bartolini C., Trastour D., "A Decision Support Tool for Optimizing Scheduling of IT Changes", 10th IFIP/IEEE Symp. On Integrated Mgmt, 2007.

[6] A. keller, J. L. Hellerstein, J. L. Wolf, K. –L. Wu, V. Krishnan. "The CHAMPS System: Change Management with Planning and Scheduling" in 9th IEEE/ IFIP Network Operations and Management Sysposium (NOMS 2004), Seoul, Korea, Abril 2004.

[7] Sauvé, J., Moura, A., Sampaio, M., Jornada, J. and Radziuk, E., "An Introductory Overview and Survey of Business–Driven IT Management", in Proceedings of the 1st IEEE / IFIP International Workshop On Business-Driven IT Management, in conjunction with NOMS 2006, Vancouver, Canada, pp. 1-10.

[8] Machiraju, V., Bartolini, C. and Casati, F., "Technologies for Business-Driven IT Management", in "Extending Web Services Technologies: the Use of Multi-Agent Approaches", edited by Cavedon, L., Maamar, Z., Martin, D. and Benatallah, B., Kluwer Academic, 2005, pp.1-28.

[9] Sauvé, J. P., Marques, Filipe Teixeira, Moura, José Antão Beltrão, Sampaio, Marcus Costa, Jornada, João, Radziuk, Eduardo Optimal Design of E-Commerce Site Infrastructure from a Business Perspective In: Hawaii International Conference on System Sciences, 2006, Waikoloa, Hawaii. Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06). Washington DC: IEEE Computer Society, 2006. v.8. p.178.3 - 178.3

[10] Bartolini, C., Sauvé, J. and Sahai, A. (eds.), "Information Technology Management from a Business Perspective. Proceedings of the 1st IEEE / IFIP International Workshop on Business-Driven IT Management (collocated with NOMS 2006), Vancouver, Canada,April 2006

[11] Abrahão, B., Almeida, V., and Almeida, J., "A Self-Adaptive SLADriven Capacity Management Model for Utility Computing", Elsevier Service, 2005.

[12] Aib, I., Sallé, M., Bartolini, C. and Boulmakoul, A., "A Business Driven Management Framework for Utility Computing Environments", HP Labs Bristol Tech. Report, 2004-171.

[13] B. Abrahão, V. Almeida, and J. Almeida, "Self-adaptive SLA-driven capacity management for internet services," in 17th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management,DSOM 2006, 2006.

[14] Brown, A.B., Keller, A., Hellerstein, J.L., "A model of configuration complexity and its application to a change management system", in: Integrated Network Management, 2005. IM 2005, pp. 631- 644.

[15]Christian Mayerl, Jens-Uwe Gaspar, Christof Momm, Sebastian Abeck, " Definition of Metric dependencies for Monitoring the impact of Quality of Services on Quality of Processes", in: BDIM – Business-Driven IT management 2007.

[16] Gilad Barash, Claudio Bartolini, Liya Wu, "Measuring and Improving the Performance of an IT Support Organization in Managing Service Incidents" in: BDIM – Business-Driven IT management 2007.

[17] J.A. Nilsson and A.U. Ranerup. Elaborate change management: Improvisational introduction of groupware in public sector. In Proceedings of the 34th Annual Hawaii International Conference on System Sciences, 2001.

A Low-Power Content-Addressable Memory (CAM) using Pipelined Search Scheme

Yibo Song, Zheng Yao, Xingguo Xiong
Department of Electrical and Computer Engineering,
University of Bridgeport, Bridgeport, CT 06604, USA

***Abstract* — A Content-Addressable Memory (CAM) is a memory that implements the lookup-table function in a single clock cycle using dedicated comparison circuitry. CAMs are especially popular in network routers for packet forwarding and packet classification, but they are also beneficial in a variety of other applications that require high-speed table lookup. The main CAM design challenge is to reduce power consumption associated with the large amount of parallel active circuitry, without sacrificing speed or memory density. In this project, a low power content-addressable memory using pipelined hierarchical search scheme is implemented. The search operation is pipelined and the match-lines are broken to several segments. Whenever the stored words fail to match in their initial segments, the search operation is discontinued for subsequent segments, hence the power is saved. The proposed scheme has been implemented in a 64×64 bit ternary CMOS CAM. The schematics of both the pipelined and non-pipelined CAMs are designed. PSPICE power simulation is used to extract the power consumption of both memory designs. The simulation results demonstrate an effective overall power reduction in the pipelined CAM compared to non-pipelined architecture for the given example memory pattern.**

Keywords: Content-Addressable Memory (CAM), low power VLSI, pipeline, power simulation

I. INTRODUCTION

A Content-Addressable memory (CAM) compares input search data against a table of stored data, and returns the address of the matching data [1]. CAMs have a single clock cycle throughput, which make them faster than other hardware- and software-based search systems. CAMs can be used in a wide variety of applications which require high search speed, such as parametric curve extraction, Hough transformation, Huffman coding/decoding, Lempel–Ziv compression, and image coding. The primary commercial application of CAMs today is to classify and forward Internet Protocol (IP) packets in network routers [1]. In networks like the internet, a message such an as e-mail or a web page is transferred by first breaking up the message into small data packets of a few hundred bytes, and then sending each data packet individually through the network. These packets are routed from the source, through the intermediate nodes of the network (called routers), and reassembled at the destination to reproduce the original message. The function of a router is to compare the destination address of a packet to all possible routes, in order to choose the appropriate route. A CAM is a good choice for implementing this lookup operation due to its fast search capability.

Fig. 1 [2] illustrates how a CAM accomplishes address lookup by implementing the routing table shown in the table. As shown in Fig. 1, the packet destination-address of 01101 is the input to the CAM. According to the table, two locations match, with the (priority) encoder choosing the upper entry and generating the match location 01, which corresponds to the most-direct route. This match location is the input address to a RAM that contains a list of output ports, as depicted in Fig. 1. A RAM read operation outputs the port designation, port B, to which the incoming packet is forwarded. We can treat the match location output of the CAM as a pointer that retrieves the associated word from the RAM. In the particular case of packet forwarding, the associated word is the designation of the output port. This CAM/RAM system is a complete implementation of an address-lookup engine for packet forwarding.

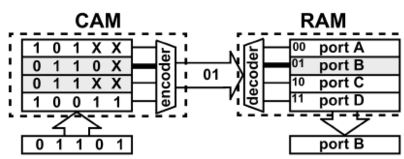


TABLE I EXAMPLE ROUTING TABLE

Entry No.	Address (Binary)	Output Port
1	101XX	*A*
2	0110X	*B*
3	011XX	*C*
4	10011	*D*

Figure 1: CAM-based implementation of an example routing table [2]

However, the speed of a CAM comes at the cost of increased silicon area and power consumption, two design parameters that designers strive to reduce. As CAM applications grow, demanding larger CAM sizes, the power problem is further exacerbated. Reducing power consumption without sacrificing speed or area is the main thread of recent research in large-capacity CAMs. Various low power CAM memory designs have been proposed [3]-[9]. In this paper, we mainly implement the concepts in [3] for a 64×64 bit ternary CMOS pipelined CAM, and compare its power consumption with the non-pipelined design. Simulation results demonstrate effective power savings in

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_71,

the low power pipelined CAM memory design for the given example memory pattern.

II. CAM Basics and Matchline Structure

A small model of CAM architecture is shown in Fig. 2 [4]. The figure shows a CAM consisting of 4 words, with each word containing 3 bits arranged horizontally (corresponding to 3 CAM cells). There is a matchline corresponding to each word feeding into matchline sense amplifiers (MLSAs), and there is a differential searchline pair corresponding to each bit of the search word. A CAM search operation begins with loading the search-data word into the search-data registers followed by precharging all matchlines high, putting them all temporarily in the match state. Next, the searchline drivers broadcast the search word onto the differential searchlines, and each CAM core cell compares its stored bit against the bit on its corresponding searchlines. Matchlines on which all bits match remain in the precharged-high state. Matchlines that have at least one bit that misses, are discharged to ground. The MLSA then detects whether its matchline has a matching condition or miss condition. Finally, the encoder maps the matchline of the matching location to its encode address.

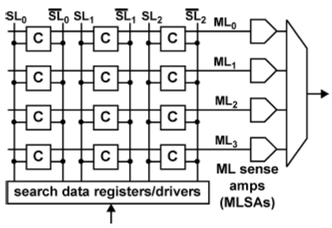


Figure 2: Simple schematic of a CAM model [4]

Binary Versus Ternary CAMs

CAMs can be divided into two categories [3]: binary CAMs and ternary CAMs (TCAMs). A binary CAM can store and search binary words (consisting of '0's and '1's). Thus, binary CAMs are suitable for applications that require only exact-match searches. A more powerful and feature-rich TCAM can store and search ternary states ('1', '0', and 'X'). The state 'X', also called 'mask' or 'don't care', can be used as a wild card entry to perform partial matching. Masking can be done both globally (in the search key) and locally (in the table entries). Figure 3 shows examples of global and local masking in TCAMs. In Figure 3(a), the search key 10110XXX matches with all the entries that fall in the following range: 10110000 to 10110111 (words located at addresses 1 and 4 in this case). It is called global masking because the last three bits of all the table entries are ignored. In Figure 3(b), word 110-XX-010 (located at address 2) will match with any of the following search keys: 110-00-010, 110-01-010, 110-10-010 and 110-11-010. The mask feature is particularly suitable for longest-prefix match searches in classless inter-domain routing (CIDR). TCAMs are also becoming popular in solving other problems such as sorting and range searching.

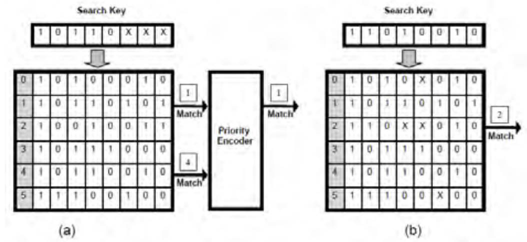


Figure 3: Search operation in a TCAM with (a) global masking and (b) local masking [5]

The Binary CAM Cell

A CAM cell serves two basic functions: bit storage (as in RAM) and bit comparison (unique to CAM). Fig. 4 shows a NOR-type CAM cell [Fig. 4(a)] and the NAND-type CAM cell [Fig. 4(b)]. The bit storage in both cases is an SRAM cell where cross-coupled inverters implement the bit-storage nodes D and $\overline{D}$. To simplify the schematic, we omit the NMOS access transistors and bitlines which are used to read and write the SRAM storage bit. Although some CAM cell implementations use small area DRAM cells, typically, CAM cells use SRAM storage. The bit comparison, which is logically equivalent to an XOR of the stored bit and the search bit is implemented in a somewhat different fashion in the NOR and the NAND cells.

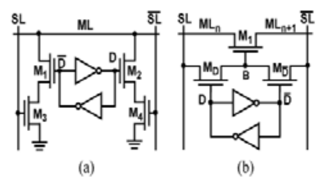
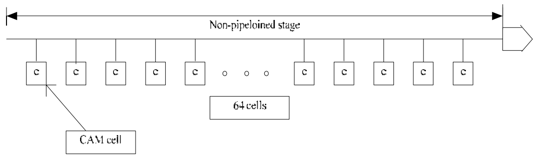


Figure 4. CAM core cells for (a) 10-T NOR-type CAM and (b) 9-T NAND-type CAM [9]

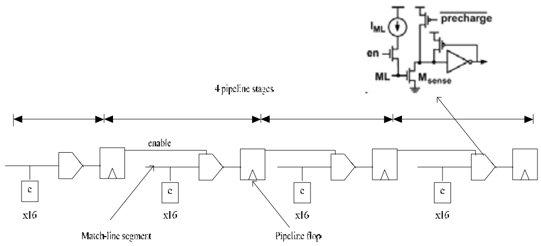
III. A low-Power Content-Addressable Memory design

This paper is mainly based on the concept in [3]. The technique is to pipeline the search operation by breaking the match-lines into several segments. Since most stored words fail to match in their first segments, the search operation is discontinued for subsequent segments, hence reducing the power.

Fig. 5 shows the non-pipelined as well as pipelined matchline architectures. Both the non-pipelined and pipelined architectures use a current-based MLSA. In this ML sensing scheme, the match-line is pre-discharged to low and a current is forced into the ML. MLs in the miss state discharge the current to ground and thus there is little increase in the ML voltage. MLs in the match state collect charge and the ML voltage increases. The NMOS transistor turns on only for MLs in the match state, which in turn flips the state of the ensuing half-latch, indicating a match.



(a). Non-pipelined match-line architecture



(b). Pipelined match-line architecture

Figure 5. Non-piplelined and pipelined match-line architecture of the CAM

Figure 6 shows the schematic design of the 64×64 Content-Addressable Memory in PSPICE. Since the overall memory is too large to fit in the design space, hierarchical design strategy is utilized. As shown in Fig. 6, the overall memory consists of 4×4 modules, with each module representing a 16×16 block. Figure 7 shows the detailed schematic design of one memory module. The ML is divided into four ML segments, each evaluated sequentially in a pipeline fashion. Each segment has 16 bits, hence the ML has totally 64 bits. The MLSA current source that provides the I_{ML} current is divided among the four segments in proportion to the number of bits in each segment. This is to guarantee the identical speed in all ML segments and to allow a fair comparison with the non-pipelined architecture. The pipelined ML operates from left to right, with each ML segment acting as an enable signal for the MLSA of the subsequent segment. Hence, only words that match a segment proceed with the search in their subsequent segments. Words that fail to match a segment do not search for their subsequent segments and hence power is saved.

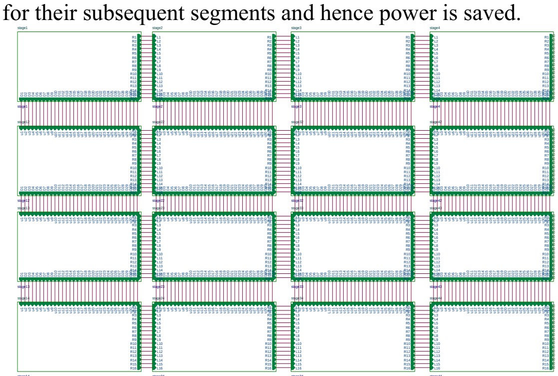


Figure 6. Schematic design of 64×64 CAM

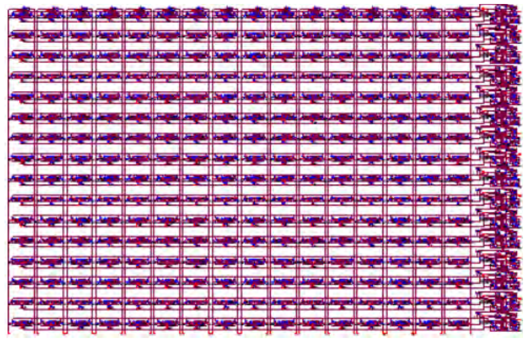


Figure 7. Schematic design of one module in 64×64 CAM

The key benefit of the pipelined architecture is that it exploits the same effect as the selective precharge scheme to reduce match-line power consumption. In typical CAM applications, such as router address look-up, only one or two words match and all other words miss. Furthermore, most words will miss on the first few bits. This fact is exploited in power reduction by allocating the first segment so that the majority of words will miss in this segment. The subsequent segments are set to be relatively larger in size in order to minimize the total amount of duplicated sensing and pipeline circuitry required.

In order to verify the power savings of the pipelined CAM memeory design, PSIPICE simulation is used to etract the power consumption of both the non-pipelined and pipelined CAM memory circuits for the given input patterns. In PSPICE power simulation, an auxiliary power measurement circuit (including a current-controlled current source $k \cdot i_{dd}$, a capacitor C and a resistor R) is added, as shown in Fig. 8.

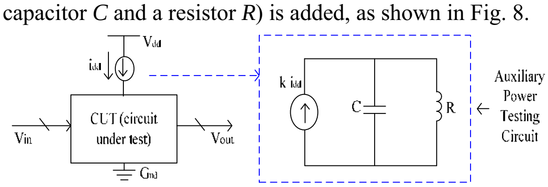


Figure 8. Power analysis using PSPICE

The average power of circuit under test (CUT) in time period T **is:**

$$P_{avg}(T) = \frac{E(t)}{T} = \frac{1}{T}\int_0^T P(t)dt = \frac{V_{dd}}{T}\int_0^T i_{dd}(t)dt \qquad (1)$$

In power testing circuit:

$$k \cdot i_{dd}(t) = \frac{dQ_c}{dt} = \frac{d(CV_C)}{dt} = C\frac{dV_C(T)}{dt} \Rightarrow V_C(T) = \frac{k}{C}\int_0^T i_{dd}(t)dt \qquad (2)$$

Compare above two, if we select: $k = (V_{dd} \cdot C) / T$

Then $\dfrac{V_{dd}}{T} = \dfrac{k}{C} \Rightarrow P_{avg}(T) = V_C(T)$ (3)

That is, the voltage in capacitor *C* is numerically equal to the average power of the CUT. Thus by measuring the voltage $V_C(T)$ in PSPICE simulation, we know the value of $P_{avg}(T)$. The power consumption of a circuit can be extracted by PSPICE simulation.

IV. Results and Discussion

We performed transient simulation to verify the function of the designed pipelined CAM memory. The following input patterns are used for all inputs:

Pattern1:
0000000000000000000000000000000000000000000000
000000000000000000
Pattern2:
0000000000000000000000000000000000000000000000
000000000000000000
Pattern3:
1111111111111111111111111111111111111111111111
111111111111111111
Pattern4:
0000000000000000000000000000000000000000000000
000000000000000000

The precharge and evaluation phases for each pattern last for 3.5ns and 3.5ns separately. That is, clock period T_{clk}=7ns. If all the CAM memory cells match the corresponding search line bits, the MLSA will detect the match condition and feed into next stage. Fig 9 shows that all the memory bits are set as 1. As pipelined ML operates from left to right, at 10.5ns-14ns all the MLs match. Each ML segment then acts as an enable signal for the MLSA of the subsequent segment. Hence, the MLSA detects the matching condition for all stage.

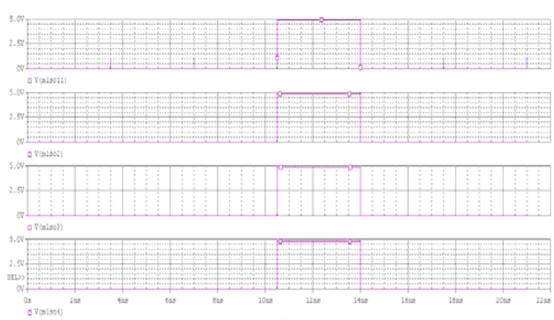


Figure 9. Simulated waveforms of CAM for all stage match

Fig 10 shows that the last stage MLSA didn't detect the matching signal in last stage. As pipelined ML operates from left to right, the first three stages are all turned on during 10.5ns-14ns which means that all the CAM memory cells in first three stages match their corresponding searching bits. Each ML segment acts as an enable signal for the MLSA of the subsequent segment. Since the memory cells in last stage didn't match their corresponding searching line bits, the last stage is not turned on. As a result, power is saved.

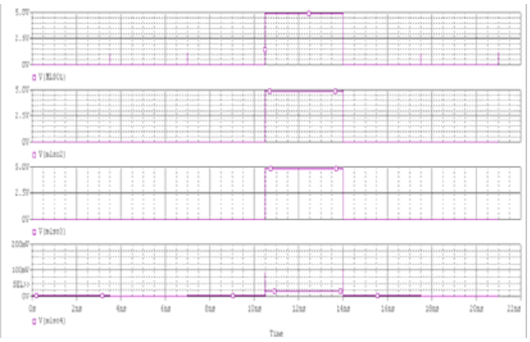


Figure 10. Simulated waveforms of CAM for last stage mismatch

Fig 11 shows that the MLSA turns off the last two stages due to a mismatch in the third stage. As pipelined ML operates from left to right, the first two stages all turn on during 10.5ns-14ns which means all the CAM memory cells in first two stages match their corresponding searching bits. Each ML segment acts as an enable signal for the MLSA of the subsequent segment. Since the third stage failed to match the searching bits, the third and fourth stages are both turned off to save the power.

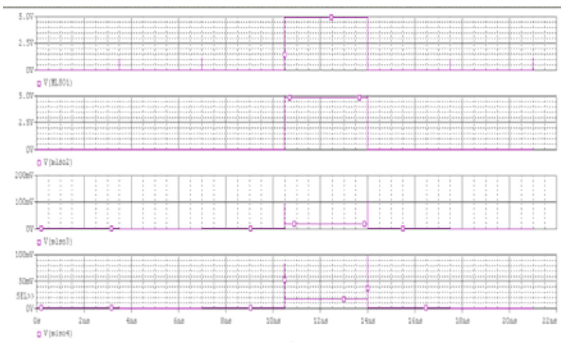


Figure 11. Simulated waveforms of CAM for third stage miss

Similarly, Fig. 12 shows that when the CAM cells mismatch the searching bits at second stage, then the MLSA turns off the subsequent stages (the third and fourth stages) and power can be saved. In Fig. 13, when the CAM cells fail to match the searching bits at the very first stage, then all the MLSA will turn off all the four stages, and a maximum power saving can be achieved. While in the non-pipelined CAM design, all the bits will be enabled in order to decide a match or a miss, which leads to a larger power consumption compared to the pipelined design.

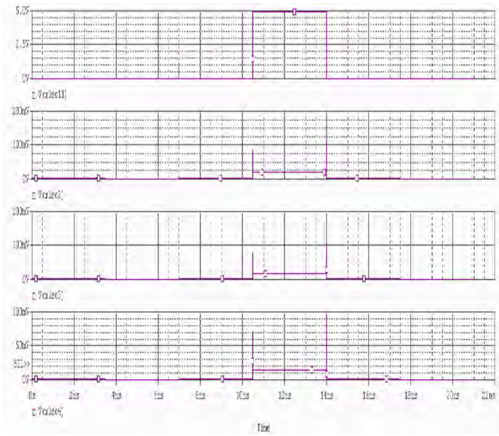


Figure 12. Simulation of CAM for second stage mismatch

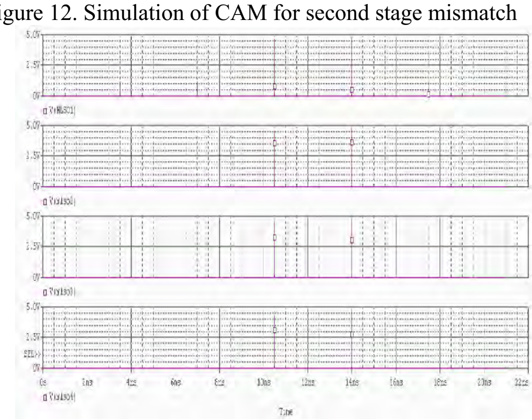


Figure 13. Simulation of CAM for all stage mismatch

In order to verify the power consumption of the pipelined CAM design, PSPICE power simulation is used to extract the power consumption of both the pipelined and non-pipelined CAM memory for the same giving patterns. Fig 14 shows the simulated power curve $V(P_{avg})$ waveform of the non-pipelined 64×64 CAM memory for the given input pattern. Based on the cursor measurement results in $V(P_{avg})$ waveform, the average power consumption of the non-pipelined CAM during time periods T_1=21ns is found to be P_{avg}=2.2041mW.

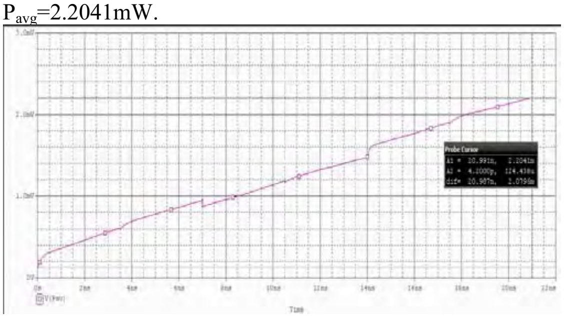


Figure 14. Power consumption for non-pipelined scheme

Fig 15 shows the simulated power curve of the 4-stage pipelined CAM design with all the 4 stages in matching condition. Based on the simulation result, the average power of the pipelined CAM during time periods T_1=21ns is found to be P_{avg}=2.7671mW. Compared with the non-pipelined CAM, the pipelined CAM does not save power in this case. This is because all the four stages are enabled due to the complete matching in all the four bits in this case. We can see that actually it spends even a little more power compared to non-pipelined CAM due to the extra pipeline registers introduced.

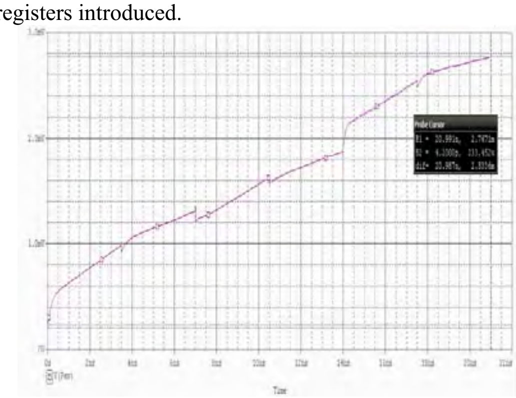


Figure 15. Power consumption for pipelined scheme (all match)

Fig 16 shows the simulated power curve of the 4-stage CAM design when the first stage is found to be a mismatch with the searching bits. In this case, all the four stages will be disabled and hence a maximum power saving is achieved. Based on the simulation results, the average power consumption of the pipelined CAM during time periods T_1=21ns is found to be P_{avg}=505.6µW, compared to the non-pipelined scheme, a significant power saving of 77.1% is achieved in this case.

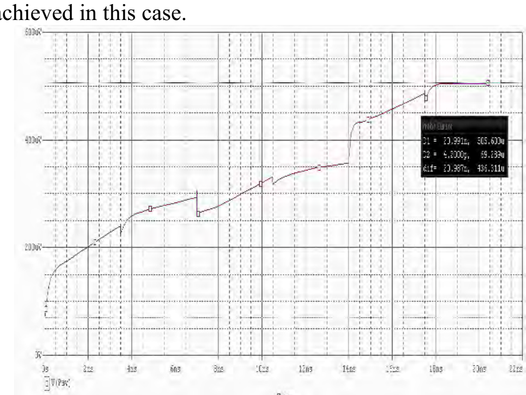


Figure 16. Power consumption for pipelined scheme (all miss)

The average power of the 4-stage pipelined CAM memory with the mismatch in second stage is also simulated. According to the simulation results, the average power of the pipelined CAM with the second stage mismatch during time period of T_1=21ns is found to be P_{avg}=561.288µW. Compared with the non-pipelined scheme, a power saving of 74% is observed. This power saving is slightly lower than the case when the first stage is mismatched. This is reasonable because the first stage is still turned on for comparison when the mismatch occurs in the second stage.

Similarly, the average power consumption of the 4-stage pipelined CAM during time period T_1=21ns when the mismatch occurs in the third stage is found to be P_{avg}=1.0417mW in PSPICE power simulation. This translates to a power saving of 52.65% compared to the non-pipelined CAM design with the same input pattern.

Further, the average power consumption of the 4-stage pipelined CAM during time period T_1=21ns when the mismatch occurs in the 4^{th} stage is found to be P_{avg}=1.45mW in PSPICE power simulation. Compared to the original non-pipelined CAM design, a power saving of 34% is observed. This leads to least power saving compared to the previous cases because three out of four stages in the pipelined CAM is activated for comparison in this case.

Compare the energy consumption of the propose architecture with the non-pipelined architecture, we can see that the addition of pipelined match-lines reduces the energy consumption, consisting of a ML energy consumption and no change in SL energy. Since most ML segments miss in the first 16-bit stage (rarely activating the subsequent 48 bits), one expects the ML energy consumption to be reduced by 74%, compared to the non-pipelined architecture. The simulation results measurement, however, the total 56% ML energy is reduction in the 64×64 pipelined content addressable memory designs.

From above simulations, we can see that generally the pipelined CAM design leads to significant power savings compared to the non-pipelined CAM design. Especially if the first stage is found to be a mismatch with the searching bits, all the four stages will be turned off and a significant power saving of 77.1% is observed for the given patterns. If the mismatch occurs in a later stage, less power saving will be resulted because more stages will be activated for comparison. If no mismatch exists, i.e. all the four stages match the searching bits, then all the four stages will be activated and no power saving can be achieved. Generally, most of the patterns lead to mismatch in one of the four stages. Hence effective power savings can be achieved for the pipelined CAM design.

V. Conclusions and Future Work

A CAM is a memory that implements the lookup-table function in a single clock cycle using dedicated comparison circuitry. CAMs are especially popular in network routers for packet forwarding and packet classification, but they are also beneficial in a variety of other applications that require high-speed table lookup. The main CAM-design challenge is to reduce power consumption associated with the large amount of parallel active circuitry, without sacrificing speed or memory density. In this paper, a low power 4-stage pipelined 64×64 CAM memory design [3] is implemented. The power savings of the pipelined match-lines is due to the fact that only a small portion of the ML segments will be activated in general searching operation, and the rest of the segments will be deactivated to save power. The schematic of the pipelined CAM memory is designed in PSPICE. Simulation results demonstrate the correct function of the design. Furthermore, PSPICE simulations are used to extract the average power consumptions of both pipelined and non-pipelined CAM memories. Simulation results show that if all the four stages mismatch, a significant power saving of 77.1% is observed for the pipelined CAM compared to the non-pipelined design. The power savings also depends on where the first mismatch occurs in the four stages. If the mismatch occurs in the early pipeline stage, more power saving will be achieved due to the factor that the rest of the stages can be turned off to save power.

Similarly, power can also be saved if the hierarchical search-lines activate only a small portion of the local searchlines (LSLs). In the future work, we will study how to save power in search-lines by using hierarchical search scheme and combine it with the pipelined scheme.

REFERENCES

[1] A. S. Tanenbaum, "*Computer Networks*," Prentice Hall, Upper Saddle River, NJ, 2003.

[2] J. M. Rabaey, A. Chandrakasan, and B. Nikoliæ, *Digital Integrated Circuits: A Design Perspective*, 2nd ed. Englewood Cliffs, NJ, Prentice-Hall, 2003.

[3] K. Pagiamtzis and A. Sheikholeslami, "Pipelined match-lines and hierarchical search-lines for low-power content-addressable memories," *Proc. IEEE Custom Integrated Circuits Conf.*, 2003, pp. 383–386.

[4] R. Sangireddy, and A. K. Somani, "High-speed IP routing with binary decision diagrams based hardware address lookup engine," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 4, pp. 513-521, May 2003.

[5] K. Etzel, "Answering IPv6 lookup challenges," Technical Article, Cypress Semiconductor Corporation, Oct. 27, 2004.

[6] A. J. McAuley, and P. Francis, "Fast routing table lookup using CAMs," Proceedings INFOCOM'93, San Francisco, CA, vol. 3, pp. 1382-1391, 1993.

[7] T.-B. Pei and C. Zukowski, "Putting routing tables in silicon," *IEEE Network Mag.*, vol. 6, pp. 42–50, Jan. 1992.

[8] L. Chisvin and R. J. Duckworth, "Content-addressable and associative memory: Alternatives to the ubiquitous RAM," *IEEE Computer*, vol. 22, pp. 51–64, July 1989.

[9] H. Miyatake, M. Tanaka, and Y. Mori, "A design for high-speed lowpower CMOS fully parallel content-addressable memory macros," *IEEE J. Solid-State Circuits*, vol. 36, pp. 956–968, June 2001.

Design and Optimization of Piezoelectric Dual-Mode Micro-Mirror

Jichao Zhong, Xingguo Xiong, Zheng Yao, Junling Hu*, Prabir Patra*
Department of Electrical and Computer Engineering,
*Department of Mechanical Engineering
University of Bridgeport, Bridgeport, CT 06604, USA

***Abstract*—MEMS micro-mirrors have been widely used in optical communication, projection display, and microscanner, etc. Various MEMS micro-mirror devices have been reported. In this paper, the design of a bulk-micromachined MEMS piezoelectric dual-mode micro-mirror is proposed. The working principle of the dual-mode micro-mirror device is analyzed. Based on the analysis, a set of optimized design parameters of the micro-mirror is suggested. Theoretical analysis shows that the proposed micro-mirror can deflect for a maximum displacement of 2.5µm in its piston mode, and tilt for a maximum angle of 4.88º in its torsional mode. The fabrication flow of the MEMS micro-mirror is also suggested. The proposed micro-mirror has the advantages of large displacement and large tilt angle due to piezoelectric activation. The PZT actuated micro-mirror (PAM) also leads to improved linearity in tilting angle and piston displacement. Further, fast response of the PZT unimorph on the applied actuation voltage ensures wider operation bandwidth of the PAM in contrast to thermal activated micro-mirrors.**

Keywords: MEMS (Microelectromechanical System), dual-mode micro-mirror, piezoelectric technology (PZT), piston micro-mirror, torsional micro-mirror.

I. Introduction

Microelectromechanical Systems (MEMS) deal with devices or systems in the size range of 1µm~1mm, which integrates both mechanical and electrical components into a single chip. MEMS merges with nanotechnology and creates the field of Nanoelectromechanical Systems (NEMS). Optical MEMS is a unique application where MEMS are used to direct, guide, filter, and, in some instances, amplify light. Well known applications of optical MEMS include optical switching and digital light projection. In optical switching applications, micro-mirrors are used to steer light from an incoming fiber optic input to a fiber optic output. In digital light projection applications, micro-mirrors, such as digital micro-mirror device (DMD) by Texas Instruments Inc., incident light is reflected by an array of torsional micro-mirrors so that each mirror pixel can control a corresponding spot on the screen to be bright or dark. In this way, vivid pictures can be projected onto the screen.

According to the working modes, MEMS micro-mirrors can be divided into two categories: piston mirrors and torsional mirrors. Piston micro-mirrors move perpendicular to the device plan, while torsional micro-mirrors rotate along a certain axis. Various piston and torsional micro-mirrors have been reported [1]-[4]. The working principles and design considerations of MEMS micromirrors have also been analyzed in details [5]-[6]. Furthermore, dual-mode micro-mirrors which can work in both piston and torsional modes have been reported [7]-[8]. Dual-mode micro-mirrors offers more flexibility in light modulation. Users can reconfigure them into either piston or torsional modes according to their individual needs.dd

MEMS micro-mirrors can be activated using various mechanisms, such as electrostatic driving, electromagnetic driving, thermal driving, piezoelectric driving, as well as shape-memory alloy driving, etc. Among them, electrostatic driving is popular due to their easy in implementation. But the maximum displacement is limited by snap-down effect in perpendicular driving. Electromagnetic driving has larger activation force and displacement, but the fabrication is more complex. Thermal driving is relatively slow in response, and the energy efficiency is not high due to thermal dissipation. Piezoelectric driving utilizes piezoelectric effect to activate MEMS micro-mirrors. It can achieve large displacement or rotation angle with high energy efficiency. Various piezoelectric micro-mirrors have been reported [9]-[11].

In this paper, a bulk-micromachined piezoelectric-driven, dual-mode MEMS micro-mirror device is proposed. The dual-mode micro-mirror can be configured to work in either piston or torsional mode according to the need. The working principle of the dual-mode micro-mirror device is analyzed. Based on the analysis, a set of optimized design parameters of the micro-mirror is suggested. Theoretical analysis shows that the proposed micro-mirror can deflect for a maximum displacement of 2.5µm in its piston mode, and tilt for a maximum angle of 4.88º in its torsional mode. The fabrication flow of the MEMS micro-mirror is also suggested. The proposed micro-mirror can achieve large displacement and large tilt angle due to piezoelectric activation. The PZT actuated micro-mirror (PAM) also has improved linearity in tilting angle and piston displacement.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_72,

II. Piezoelectric Activation

Piezoelectric materials are used for both sensing and actuation purposes. It was observed that certain materials generate an electric charge (or voltage) when they are under mechanical stress. This is known as the direct effect of piezoelectricity. Alternately, the same materials would be able to produce a mechanical deformation (or force) when an electric field is applied to them. This is called the inverse effect of piezoelectricity.

Piezoelectric materials are crystals. The microscopic origin of piezoelectricity is the displacement of ionic charges within a crystal, leading to the polarization and electric field. A stress (tensile or compressive) applied to a piezoelectric crystal will alter the spacing between centers of positive and negative charge sites in each domain cell; this leads to a net polarization manifested as open circuit voltages measurable at the crystal surface. Compressive and tensile stresses will generate electric fields and hence voltages of opposite polarity.

Inversely, an external electric field will exert a force between the centers of positive and negative charges, leading to an elastic strain and changes of dimensions depending on the field polarity. Not all naturally occurring or synthesized crystals exhibit piezoelectricity. Crystals can be classified into 32 groups according to crystal symmetry. Centrosymmetric crystal structures are crystals that are symmetric along all axes through the center of the crystal. These crystals occupy 11 out of 32 possible groups and are non-piezoelectric materials because the positive and negative charge sites will not be spatially separated under stress. Out of 21 non-centrosymmetric groups, 20 are piezoelectric crystals. Piezoelectric effects are strongly orientation dependent. The notation conventions for crystal orientations in the context of piezoelectric polarization are discussed first. A piezoelectric material needs to be poled in a particular direction to provide a strong piezoelectric effect, although some materials exhibit natural or spontaneous polarization. The direction of positive polarization is customarily chosen to coincide with the Z-axis of a rectangular system of crystallographic axes X, Y, and Z. Alternatively, the normal stress components along axes X, Y, and Z are denoted by subscripts 1, 2 and 3, respectively, as shown in Figure 1. As such, the poling axis always coincides with axis-3. Shear stress and strain components about these axes are denoted by sub-scripts 4, 5 and 6 respectively.

In a piezoelectric crystal, the constitutive equation that relates electrical polarization (D) and applied mechanical stress (T) is [12]:

$$D = dT + \varepsilon E \tag{1}$$

where d is the piezoelectric coefficient matrix, ε is the electrical permittivity matrix, and E is the electrical field. Here, an electric field is applied in conjunction with the mechanical stress to provide more generality. The electrical polarization is contributed by two parts – one stemming from electrical biasing and one from mechanical loading.

If no electric field is present (i.e., E=0), then the second term on the right-hand side of the above equation can be eliminated.

The general constitutive equation can be written in the full matrix form:

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}$$

the terms T_1 through T_3 are normal stresses along axes 1, 2 and 3, whereas T_4 through T_6 are shear stresses. The units of electrical displacement (D_i) , stress (T_i), permittivity (ε_i), and electrical field (E_j) are C/m^2, N/m^2, F/m^2, and V/m, respectively. The unit of the piezoelectric constant d_{ij} is the unit of electric displacement divided by the unit of the stress, namely:

$$[d_{ij}] = \frac{[D]}{[T]} = \frac{[\varepsilon][E]}{[T]} = \frac{\frac{F}{m}\frac{V}{m}}{\frac{N}{m^2}} = \frac{Columb}{N}$$

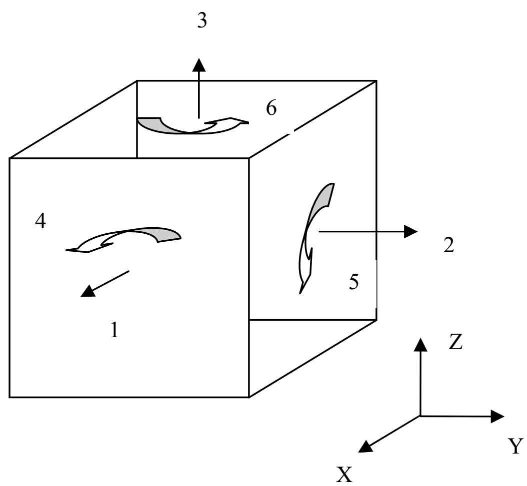


Figure 1: Schematic illustration of a piezoelectric crystal in a Cartesian coordinate system [12]

The inverse effect of piezoelectricity can be similarly described by a matrix-form constitutive equation. In this case, the total strain is related to both the applied electric field and any mechanical stress [12], according to

$$s = ST + dE$$

Where s is the strain vector and S is the compliance matrix.

This equation can be expanded to a full matrix form [12]:

$$\begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} & S_{15} & S_{16} \\ S_{21} & S_{22} & S_{23} & S_{24} & S_{25} & S_{26} \\ S_{31} & S_{32} & S_{33} & S_{34} & S_{35} & S_{36} \\ S_{41} & S_{42} & S_{43} & S_{44} & S_{45} & S_{46} \\ S_{51} & S_{52} & S_{53} & S_{54} & S_{55} & S_{56} \\ S_{61} & S_{62} & S_{63} & S_{64} & S_{65} & S_{66} \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \end{bmatrix} + \begin{pmatrix} d_{11} & d_{21} & d_{31} \\ d_{12} & d_{22} & d_{32} \\ d_{13} & d_{23} & d_{33} \\ d_{14} & d_{24} & d_{34} \\ d_{15} & d_{25} & d_{35} \\ d_{16} & d_{26} & d_{36} \end{pmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}$$

if there is no mechanical stress present ($T_{i,i=1,6}=0$), the strain is related to the electric field by [12]

$$\begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \end{bmatrix} = \begin{pmatrix} d_{11} & d_{21} & d_{31} \\ d_{12} & d_{22} & d_{32} \\ d_{13} & d_{23} & d_{33} \\ d_{14} & d_{24} & d_{34} \\ d_{15} & d_{25} & d_{35} \\ d_{16} & d_{26} & d_{36} \end{pmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}$$

Note that, for any given piezoelectric material, the d_{ij} components connecting the strain and the applied field in the inverse effect are identical to the d_{ij} connecting the polarization and the stress in the direct effect.

The electromechanical coupling coefficient k is a measure of how much energy is transferred from electrical to mechanical energy, or vice versa, during the actuation process [12]:

$$k^2 = \frac{energy\_converted}{input\_energy}$$

this relation holds true for both mechanical-to-electrical and electrical-to-mechanical energy conversion. The magnitude of k is a function of not only the material, but also the geometries of the sample and its oscillation mode.

III. Design and Optimization of the Micro-mirror

The schematic diagram of the proposed dual-mode piezoelectric micro-mirror is shown in Figure 1. ANSYS FEM simulation is used to verify the device function, as shown in Figure 2. As shown in Figure 1 and 2, a MEMS micro-mirror is connected to four torsional beams which extrudes from two torsional axes. On two of the four torsional beams, PZT piezoelectric materials are deposited to activate the torsional beams. When electrical driving voltages are applied to the PZT actuators on torsional beams, the PZT will expand or shrink, which generates corresponding torque to activate the micromirror to rotate/tilt along the torsional axes. In this way, the micro-mirror works in torsional mode. The two torsional axes are again connected to the outer frame, which are anchored to the substrate through four flexible piston beams. PZT actuators are also deposited on top of the four piston beams. When the electrical driving voltages are simultaneously applied to the four PZT actuators on the four piston beams, the PZT actuators will expand or shrink simultaneously, which in turn activates the frame (and the mirror) to move perpendicular to the device plane. Hence, the micro-mirror works in piston mode. By applying electrical driving voltages to the PZT actuators on either the piston or torsional beams, the micro-mirror can work in either piston or torsional mode. Hence this is a piezoelectric activated dual-mode micro-mirror design.

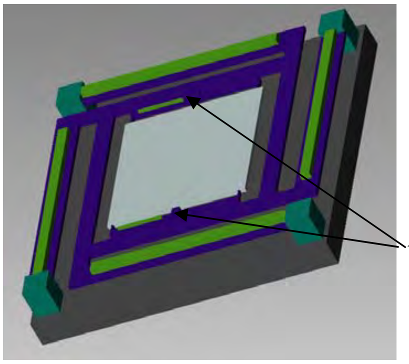


Figure 2. Schematic diagram of the dual-mode piezoelectric micro-mirror

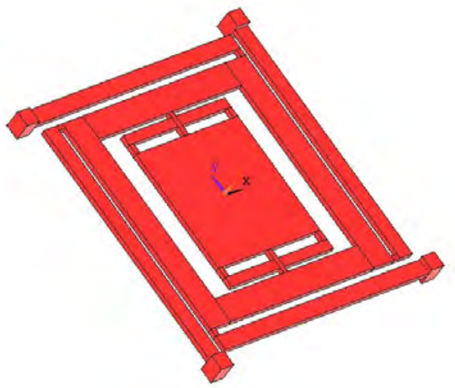


Fig 3: ANSYS model of the dual-mode micro-mirror

In this mirror design, PZT actuated micro-mirror (PAM) can provide improved linearity tilting angle and piston displacement. In addition, its displacement in piston mode is not limited by pull-in phenomenon as in perpendicular electrostatic actuation. Fast response of the PZT unimorph membrane on the applied actuation voltage ensures wider operation bandwidth of the PAM in contrast to the thermal activation. In this work, PZT-4 is chosen as piezoelectric material and polysilicon is used as substrate material for PZT actuators.

In this piezoelectric actuated micro-mirror, cantilever piezoelectric actuator model is used to analyze the device behavior. Piezoelectric actuators are often used in conjunction with cantilevers or membranes for sensing and actuation purpose. General models for such piezoelectric actuators are rather complex. Accurate analysis often involves finite element modeling. For limited cases, such as a cantilever actuator with two layers, analytical solution has been successfully achieved. Based on the theoretical analysis, a set of optimized design parameters of the micro-mirror is achieved, as shown in Table 1.

Table 1. Optimized design parameters of the dual-mode micro-mirror

Outer actuator beam	Length	100 μm
	Width	6 μm
	Thickness	2 μm
Main frame (square)	Length of side	90 μm
	Width of side	10 μm
	Thickness	4 μm
Torsional axis	Length	10/20 μm
	Width	2 μm
	Thickness	2 μm
Inner actuator beam	Length	24 μm
	Width	4 μm
	Thickness	2 μm
Gap between the outer actuator beam and the main frame	2.5 μm	
Gap between the inner actuator beam and the mirror	2.5 μm	
Gap between the mirror and the substrate	3 μm	

The deflection of a two-layer piezoelectric structure can be derived based on above model. Consider a cantilever with two layers, one elastic and one piezoelectric, joined along one side. Assume the two layers have the same length. The beam bends into an arc when the piezoelectric layer is subjected to a longitudinal strain, s_{long}. The radius of curvature can be found by [12]

$$\frac{1}{r}=\frac{2\left|s_{long}\right|\left(t_p+t_e\right)\left(A_pE_pA_eE_e\right)}{4\left(E_pI_p+E_eI_e\right)\left(A_pE_p+A_eE_e\right)+\left(t_e+t_p\right)^2A_eE_eA_pE_p}$$

In which A_p and A_e are the cross-sectional area of the piezoelectric and the elastic layer; E_p and E_e are Young's modulus of the piezoelectric layer and Young's modulus of the elastic layer; and t_p and t_e are the thickness of the piezoelectric layer and the elastic layer.

The amount of force achievable at the free end of a piezoelectric bimorph actuator equals the force required to restore the tip of the actuator to its initial un-deformed state. Since the displacement is linearly related to force according to

$$\delta(L)=F/k$$

the force can be expressed as

$$F=\delta(x=L)\cdot k$$

In this work, PZT-4 is chosen as the piezoelectric material. For PZT-4,

$$s=\begin{bmatrix}12.3 & -4.05 & -5.31 & 0 & 0 & 0\\ -4.05 & 12.3 & -5.31 & 0 & 0 & 0\\ -5.31 & -5.31 & 15.5 & 0 & 0 & 0\\ 0 & 0 & 0 & 39 & 0 & 0\\ 0 & 0 & 0 & 0 & 39 & 0\\ 0 & 0 & 0 & 0 & 0 & 39\end{bmatrix}\times 10^{-12}m^2/N$$

$$d=\begin{bmatrix}0 & 0 & 0 & 0 & 496 & 0\\ 0 & 0 & 0 & 496 & 0 & 0\\ -123 & -123 & 289 & 0 & 0 & 0\end{bmatrix}$$

$$\varepsilon_r=\begin{bmatrix}1475 & 0 & 0\\ 0 & 1475 & 0\\ 0 & 0 & 1300\end{bmatrix}$$

According to the structure we used for the actuator, we should choose d_{31} as the piezoelectric parameter. So the displacement can be calculated as below:

$$\delta(x)=\frac{x^2d_{31}\left(t_p+t_e\right)\left(A_pE_pA_eE_e\right)}{4\left(E_pI_p+E_eI_e\right)\left(A_pE_p+A_eE_e\right)+\left(t_e+t_p\right)^2A_eE_eA_pE_p}$$
$$=1.05\times10^{-6}\times V(\mu m)$$

Under small-deflection approximation, the dual-mode micro-mirror can be treated as simplified spring-mass model. In the piston mode, the four piston beams can be modeled as double-clamped beams. The spring constant of one double-clamped beam can be calculated as

$$K_b=\frac{12EI_b}{L_b^3}$$

Hence in our design, the displacement at the end of the piston beam can be calculated as

$$\delta(x)=\frac{1.05\times10^{-6}\times V}{4}=0.25V(\mu m)$$

The designed driving voltage of this micro-mirror is 10V, so the maximum displacement of the micro-mirror in piston mode is **2.5 μm.**

When the driving electrical voltage is applied to the PZT actuators in the inner torsional beams, the PZT actuators will generate corresponding torque on the torsional beams, hence the micro-mirror will tilt along the torsional axes. As a result, the dual-mode micro-mirror works in torsional mode. Assume the voltage is 50V, and the length of the torsional axis is $2\mu m(h)\times2\mu m(b)\times4\mu m(l)$, based on the above design parameters, the displacement of the actuator is calculated to be

$$\delta(x)=0.975(\mu m)$$

The tilt angle θ is calculated to be

$$\theta=2.32^{\circ}$$

we can optimize the mirror design on inner actuator, the actuator will counteract the torque of the torsional axis, and the tilt angle of the torsional axis must follow this equation:

$$\varphi = \frac{Tl}{G\beta hb^3}$$

where T is torque; l is the length of the torsional axis, G is shear modulus (for polysilicon, G=69GPa), β is a constant, h is the height of torsional beam, and b is the width of the torsional beam. Now, the size of the torsional axis can be redesigned as 2μm(h)×2μm(b)×4μm(l). For $\beta = 0.141$, under the same torque, the displacement of the micro-mirror is found to be in linear relationship with the torsional driving voltage:

$$\delta(x) = 0.041V(\mu m)$$

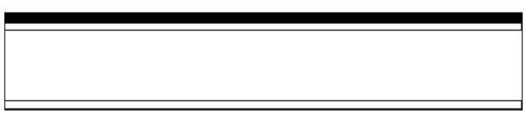
For a torsional driving voltage of 50V, the corresponding displacement and the torsional angle of the micromirror are found to be

$$\delta(x) = 2.05(\mu m)$$

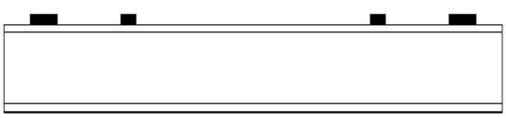
$$\theta = 4.88^{\circ}$$

IV. Fabrication Flow

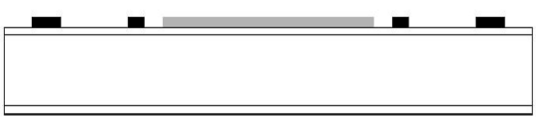
A fabrication process of the above dual-mode micro-mirror is illustrated in Figure 4**.** Starting from a bare silicon wafer, we coat the top side of the silicon wafer with 2μm thickness LPCVD low stress silicon nitride. In order to pattern the PZT layer, a thin layer of 150 nm-thick platinum layer is pre-deposited at the bottom electrode layer. Then a 0.2μm-chick PZT-4 layer is deposited using sol-gel method. On the top of the PZT-4 layer, we deposited a 150 nm-thick ruthenium oxide layer. This forms the PZT piezoelectric actuator structures for the micro-mirror.



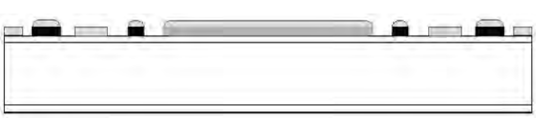
(a) Silicon wafer coated with Si_3N_4 and PZT film



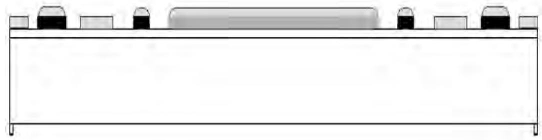
(b) Patterning of the PZT layer is delineated by self-aligned process using Cl/O-based plasma etching.



(c) The aluminum mirror surface is patterned by lift-off process.



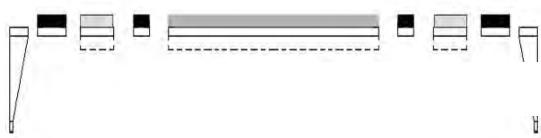
(d) Thin hard mask layer is deposited and patterned.



(e) The silicon nitride at the backside of the wafer is patterned by double side aligning and subsequent RIE to define the etch window for silicon bulk etching.

(f) The silicon under the suspended microstructures is removed by anisotropic etching in aqueous KOH. The movable microstructure can be released to be free-standing.

(g) Suspended structures including the mirror plate are released by RIE from the front side of the wafer using the patterned metal film at the step (d) as the masking material.



(h) Finally, released mirror are diced into individual chips after removing the hard mask metal

Figure 4. Fabrication flow of the piezoelectric dual-mode micromirror

V. Conclusions and Future Work

In this paper, a piezoelectric actuated dual-mode micro-mirror design is proposed. The working principle of the micro-mirror is analyzed. The micro-mirror is actuated by four PZT actuator beams in piston mode. The inner set of two PZT actuators is used to tilt the micro-mirror in torsional mode. Each set of actuators can work independently. Based on the analysis, a set of optimized design parameters of the micro-mirror is achieved. The two modes (piston and torsional) can be use for different situations as needed. Theoretical analysis shows that the proposed micro-mirror can deflect for a maximum displacement of 2.5μm in its piston mode, and tilt for a maximum angle of 4.88º in its torsional mode. The fabrication flow of the MEMS micro-mirror is also suggested. The proposed micro-mirror can achieve large displacement and large tilt angle due to piezoelectric activation. The PZT actuated micro-mirror (PAM) also has improved linearity in tilting angle and piston displacement.

In the future work, we will further improve the structure design, so that the driving voltage can be reduced from current 10V in piston mode to around 5V. In this way, it can be compatible with the power supply of on-chip CMOS circuitry. By reducing the required driving voltage, the power consumption of the micro-mirror can also be further reduced.

References

[1] G. Zhou, V.J. Logeeswaran, F.E.H. Tay, F.S. Chau, "Digital deflection micro-mirrors for adaptive optics applications", *12th International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers'03),* Vol. 1, June 8-12, 2003, pp. a591-a594.

[2] K. Jia, S. Pal, H. Xie, "An electrothermal tip–tilt–piston micromirror based on folded dual S-shaped bimorphs", *Journal of Microelectromechanical Systems*, Vol. 18, Issue 5, Oct. 2009, pp. 1004-1015.

[3] G. Vdovin, S. Middelhoek, L. Sarro, "Deformable mirror display with continuous reflecting surface micromachined in silicon", *IEEE Proceedings of Micro Electro Mechanical Systems (MEMS'95)*, Jan. 29 - Feb. 2, 1995, pp. 61–65.

[4] K.B. Lee, L. Lin, "A vertically-supported two-axial torsional micro-mirror", *Proceedings of 17th IEEE International Conference on Micro Electro Mechanical Systems (MEMS'04)*, 2004, pp. 41-44.

[5] O. Degani, Y. Nemirovsky, "Design considerations of rectangular electrostatic torsion actuation based on new analytical pull-in expression", *Journal of Microelectromechanical Systems,* Vol. 11, pp. 20–26, 2002.

[6] X. M. Zhang, F. S. Chau, C. Quan, Y. L. Lam, A. Q. Liu, "A study of the static characteristics of a torsional micro-mirror", *Sensors and Actuators A: Physical*, Vol. 90, 2001, pp. 73–81.

[7] U. Krishnamoorth, K. Li, K. Yu, D. Lee, J. P. Heritage, O. Solgaard, "Dual-mode micro-mirrors for optical phased array applications", *Sensors and Actuators A: Physical*, Vol. 97-98, 2002, pp. 21-26.

[8] X. Xiong, T. Dallas, S. Gangopadhyay, J. Berg, T. Henryk, "Design and Simulation of Aluminum Bi-functional Spatial Light Modulator", *Proc. of the* 48th *IEEE International Midwest Symposium on Circuits & Systems (MWSCAS'05)*, Cincinnati, Ohio, USA, Aug 7-10, 2005, pp. 159-162.

[9] Y. Yee, J. U Bu, M. Ha, J. Choi, H. Oh, S. Lee, and H. Nm, "Fabrication and characterization of a PZT actuated micro-mirror with two-axis rotational motion for free space optics", *Proc. of IEEE MEMS Workshop 2001*, pp. 317–320.

[10] H. Nam, Y. Kim, S. Cho, Y. Yee and J. Bu, "Low voltage PZT actuated tilting micro-mirror with hinge structure", *Conference Digest of 2002 IEEE/LEOS International Conference on Optical MEMs*, 2002, pp. 89-90.

[11] S. J. Kim, Y. H. Cho, H. J. Nam, J. U. BU, "Piezoelectrically pushed rotational micro-mirrors for wide-angle optical switch applications", *IEEE The Sixteenth Annual International Conference on Micro Electro Mechanical Systems (MEMS'03)*, Kyoto, Japan, Jan. 19-23, 2003, pp. 263-266.

[12] C. Liu, "Foundation of MEMS", Chapter 7: "Piezoelectric Sensing and Actuation", Prentice Hall, Sept. 5, 2005, ISBN: 0131472860.

Device for Measurement of Clamped Joints Friction Torque

Ing. Eva Batešková, PhD., Ing. Jozef Maščenik,PhD., Ing. Jozef Haľko, PhD., Ing. Štefan Gašpár
Faculty of manufacturing Technologies, TU Košice with seat in Prešov, Štúrova 31,
08001 Prešov, Slovak republic
eva.bateskova@tuke.sk, jozef.mascenik@tuke.sk, jozef.halko@tuke.sk, stefan.gaspar@tuke.sk

***Abstract* - This contribution deals with a proposal of device for measurement of clamped joints. In the second chapter, we propose a model of clamped joint. In the experimental part we measured values of the friction torque depending on twisting moment, and in the conclusion we measured shaft surface abrasivity.**

INTRODUCTION

One of the basic construction elements for construction production either in engineering or building and construction industries are clamped joint claws. From the point of view of working life and criteria on the clamped joint jaws, the main importance belongs to stating the correct clamping power, as well as stating the correct friction torque in the joint. If there is a small clamping power on this joint, it can cause an insufficient friction torque and subsequently damage the clamped joint and thus damage the whole mechanism. On the contrary, high clamping power can cause a big deformation either on the hub or clamped joint claws, which can lead to cracks in the claws or disabling the mutual re-assembling and disassembling of the hub and claws [1].

I. BASIC REQUIREMENTS ON JOINTS

In appliances, there are two types of binding between components and knots – movable and non-movable. The type of movable binding results from the required function and kinematic scheme. The non-movable bindings ensure a constant position of components and knots. The non-movable bindings are obtained by linking the components and knots.

The binding elements and methods are to ensure sufficient resistance and required constant position of the components and knots. In precise mechanics, a higher number of binding methods is used.

Reasons:

- Power and deformation ratios,
- Requirements on simple assembly and disassembly,
- High requirements on precision,
- Simultaneous use of heterogeneous materials in one knot,
- Requirements on outward form [2].

II. PROPOSAL OF 3D DEVICE FOR FRICTION JOINT MEASUREMENT

Fig. 1 shows the proposed device solution. The device consists of frame 1, to which bearing shell sliding surface 14 is screwed. Holder 8 is welded to the frame and there are pegs 11 screwed in it. Via screw 4, through collar 13, dynamometer 5, peg holder and peg, the clamping power is transferred on claws 3. Weight 12 slides axially along lever 6. Shaft 2 is from both sides placed in bearings 9, which are in shell 10. Screw 7 serves to ensure the shaft position.

<u>Measuring procedure:</u>
By the screw rotation, clamping of the shaft by means of claws is generated. From the dynamometer, the clamping power is read. By means of the weight, it slides axially in the direction from the shaft axis along the lever until the clamped joint is broken. From the weight mass and the distance from the shaft, twisting moment of the clamped joint breach is calculated. The friction torque is determined from the condition for transfer of the twisting moment. The disadvantage of this device is measurement scope restriction by the lever length and relatively complicated construction, which, however, eliminates the shortcomings from the two previous proposals.

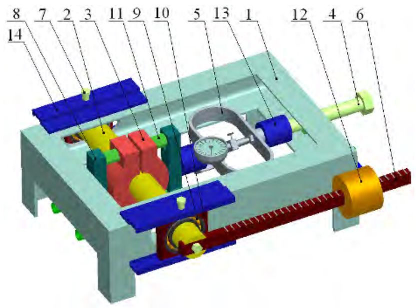


Fig. 1 3D device model

II.I PROCESS OF DEVICE SET-UP

Firstly, 8 holes were drilled into the construction frame, into which coils M8 x 1,25 were cut. Into these holes, sliding surfaces of bearing shells were screwed by means of screws with fixed head and internal hexagon. The bearings were inserted into shells and ensured by snap rings 68.

The pegs were screwed into the peg holders either into the external holes for big claws or into the internal

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_73, © Springer Science+Business Media B.V. 2010

holes for small claws. The holder with the sliding surface and collar were inserted into the tubes with the internal diameter of $\phi 12\,mm$ on the construction frame. The claws were put on the pegs according to the peg distribution. The shaft was forced into the bearings and subsequently ensured against drop-out by snap rings 40. The lever was screwed by a screw with a washer. Afterwards, a locking element against lever fall was screwed into the frame. The weight was put on the lever. The dynamometer was screwed into the collar on the peg holder and on the opposite side of the dynamometer the collar with bearing were screwed. The frame was supplemented by screw M24 x 2, which presses on the dynamometer collar and serves for educing the axial power. Finally, the entire construction was fixed to the work table.

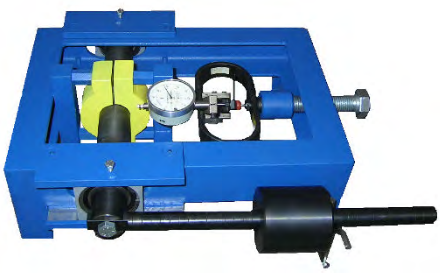


Fig. 2 Complete clamped joint device

III. EXPERIMENTAL MEASUREMENTS

The sample for this measurement is a shaft with the diametre of $\phi 40\,mm$ and one type of claws :

- For equal pressure distribution on the claw hub perimeter with external diametre of $\phi 90\,mm$

<u>Measuring method:</u>

Every measurement began by clamping the clamped joint by power of 2 kN by a screw. On the lever, the weight was set on 33 divisions, whereas every division corresponds with a different twisting moment and afterwards the screw was slowly loosened until the clamped joint was broken. The breach occurs if the lever moves. When the lever moves, we read the number of divisions from the dynamometer. Gradually, the weight moves towards the rotation axis by two division up to 9 divisions.

Subsequently, the measurement was performed from 9 divisions to 33 divisions. After these measurements, we carried out measurement of the surface abrasiveness by device Mitutoyo SJ 301, and consequently soft regrinding of claws by abrasive-coated paper and re-measurement.

Fig. 3 Device for measurement of surface abrasiveness Mitutoyo SJ 301

<u>Measurement of equal pressure distribution on hub perimeter:</u>

For this measurement, we used big claws of the clamped joint, which were put on the pegs and subsequently tightened by the screw to 2000 N. The measured values of the clamped joint breach, as well as the calculated clamping power is in chart 1. Conversion of the number of dynamometer divisions to power F was done according to the equation (1)

$$F = 2000/552{,}1.n_{divisions} \qquad (1)$$

TABLE I
MEASUREMENT ON CLAWS WITH EXTERNAL DIAMETER OF Ø 90 MM

Number of lever divisions	33	31	29	27	25
Number of dynamometer divisions	155	145	131	122	120
F [N]	561,5	525,3	474,6	441,9	434,7
Number of lever divisions	19	17	15	13	11
Number of dynamometer divisions	71	56	43	40	32
F [N]	257,2	202,9	155,8	144,9	115,9
Number of lever divisions	13	15	17	19	21
Number of dynamometer divisions	55	55	78	80	81
F [N]	199,2	199,2	282,6	289,8	293,4
Number of lever divisions	27	29	31	33	
Number of dynamometer divisions	101	105	109	110	
F [N]	365,9	380,4	394,9	398,5	

The calculation of the friction torque for equal pressure distribution was done according to equation (2)

$$M_T = F_N.f'.d = F_N.\frac{\pi}{2}.f.d \qquad (2)$$

where f is chosen from 0,15 to 0,2 , for our measurement we used the value .

TABLE II
CHART OF CALCULATED FRICTION TORQUES FOR CLAWS OF Ø 90 MM

F [N]	561,5	525,3	474,6	441,9	434,7
M_T [N.m]	5,2920	4,9508	4,4730	4,1648	4,0970
F [N]	257,2	202,9	155,8	144,9	115,9
M_T [N.m]	2,4241	1,9123	1,4684	1,3657	1,0923
F [N]	199,2	199,2	282,6	289,8	293,4
M_T [N.m]	1,8774	1,8774	2,6634	2,7313	2,7652
F [N]	365,9	380,4	394,9	398,5	
M_T [N.m]	3,4485	3,5852	3,7218	3,7558	

Fig. 4 shows the friction torque dependence on twisting moment, which results in the fact that the measurement on the device was with minor errors during the first measurement, when the weight was moved from the end of the lever towards the axis of rotation. During the measurement from the axis, a bigger dispersion arose.

Fig. 4 Friction torque dependence on twisting moment

<u>Surface abrasivity measurement:</u>

Fig. 5 shows the recording of measuring abrasivity of the surface of shaft with diameter of 50 mm.

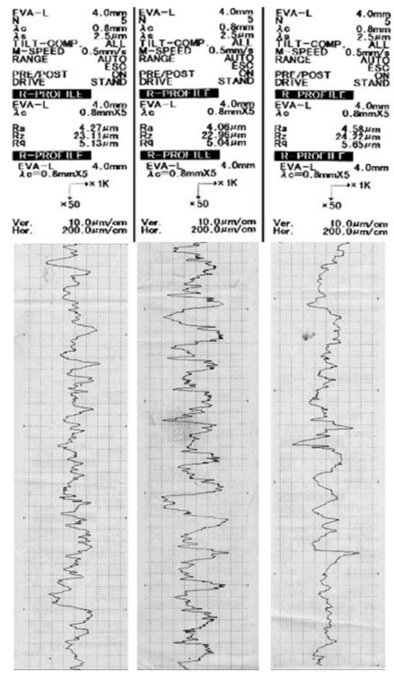


Fig. 5 Measured surface abrasivity

CONCLUSION

The calculated values of the friction torque are distorted. They depend on many factors, among which there are e.g.: precision of clamped joint make-out; surface abrasivity; clamping power; conditions of measurement; number of measurement etc. This causes the fact that when more measurements are carried out, the value of friction torque varies, which causes the dispersion in measurements. Finally, for demanding applications, it is suitable to recommend modern hydraulic tightening machines for their higher precision, and precise digital torque wrenches for clamped joint breach. This contribution forms a part of the solution of the grant task VEGA 1/4156/07 and KEGA 3/6279/08.

REFERENCES

[1] MEDVECKÝ, Š.: Základy konštruovania, EDIS, Źilina, 1999.

[2] PAVLENKO, S. – HAĽKO, J.: Časti strojov I. FVT Prešov, 2006, ISBN 80-8073-100-4.

[3] ROVŇÁKOVÁ, Sylvia - LÍŠKA, Ondrej: Predchádzanie kolíziám pri automatizovaných pracoviskách využitím modelovania a simulačných programov, Prešov, TU Košice, FVT, 2008. s. 339-342. ISBN 978-80-553-0044-3.

Self-Optimization for Dynamic Scheduling in Manufacturing Systems

Ana Madureira, Ivo Pereira
Institute of Engineering – Polytechnic of Porto, Portugal

Abstract — **Scheduling is a critical function that is present throughout many industries and applications. A great need exists for developing scheduling approaches that can be applied to a number of different scheduling problems with significant impact on performance of business organizations. A challenge is emerging in the design of scheduling support systems for manufacturing environments where dynamic adaptation and optimization become increasingly important. In this paper, we describe a Self-Optimizing Mechanism for Scheduling System through Nature Inspired Optimization Techniques (NIT).**

Index Terms— Self-Optimization, Multi-agent Learning, Autonomic Computing, Multi-Agent Systems, Nature Inspired Optimization Techniques.

I. Introduction

Scheduling is an important aspect of automation in manufacturing systems. Most of scheduling domains are characterized by a great amount of uncertainty that leads to significant system dynamics [2]. Such dynamic scheduling is receiving increased attention amongst both researchers and practitioners [2][7-9][11]. However, scheduling is still having difficulties in real world situations and, hence, human intervention is required to maintain real-time adaptation and optimization.

A challenge is emerging in the design of scheduling support systems for manufacturing environments where dynamic adaptation and optimization become increasingly important incorporating expert's knowledge. Despite the numerous advances in scheduling and MAS [6][10], most development systems still require that users have substantial knowledge of procedural-programming techniques as well as the specific computer system at hand.

At this manufacturing scheduling scenario, self-optimizing arise as the ability of the agent to monitor its state and performance, and proactively tune itself to respond to environmental stimuli.

Learning can be considered as the process of adapting behavior in response to events in the environment. Without learning, a system is limited by the ability of its designer to foresee all situations that might occur. Multi-agent Learning is the intersection of Multi-agent Systems and Machine Learning (ML), two subfields of Artificial Intelligence. Traditional ML typically involves a single agent that is trying to maximize some utility function without any knowledge. Traditional ML tasks include function approximation, classification, and problem-solving performance improvement given empirical data. MAS learning includes any situation in which an agent learns to interact with other agents [4].

This paper envisage the proposal of a Self-Optimizing mechanism for a Cooperative Scheduling System considering that AutoDynAgents [8] must be able to perform scheduling in highly dynamic environments where there is incomplete information and changes often occur; modify previously formed schedules considering recent dynamic information, minimizing the disruption of earlier schedules and still aiming for the most effective possible use of resources and achievement of goals and provide flexibility to react robustly to any disruption in an efficient and timely manner.

The remaining sections are organized as follows: Section 2 summarizes some related work on dynamic scheduling through MAS and Bio-Inspired Techniques (BIT). Some aspects of MAS and BIT are discussed on section 3 e 4. In section 5 the AutoDynAgents System is presented. Section 6 describes the Self-Optimizing Mechanism for Scheduling System through Nature Inspired Optimization Techniques. Finally, the paper presents some conclusions and puts forward some ideas for future work.

II. Dynamic Scheduling

Scheduling problems arise in a diverse set of domains, ranging from manufacturing to hospitals settings, transports, computer and space environments, amongst others. Most of these domains are characterized by a great amount of uncertainty that leads to significant system dynamism. Such dynamic scheduling is receiving increased attention amongst both researchers and practitioners.

Dynamic changes of a problem could arise from new user requirements and the evolution of the external environment. In a more general view, dynamic problem changes can be seen as a set of constraint insertions and cancellations.

For these dynamic optimization problems environments, that are often impossible to avoid in practice, the objective of the optimization algorithm is no longer to simply locate the

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_74,

global optimum solution, but to continuously track the optimum in dynamic environments, or to find a robust solution that operates optimally in the presence of perturbations [2][9]. In spite of all the previous trials, the scheduling problem is still known to be NP-complete, even for static environments. This fact poses serious challenges to conventional algorithms and incites researchers to explore new directions [9-11] and Multi-Agent technology has been considered an important approach for developing industrial distributed systems.

III. MULTI-AGENT SYSTEMS

Multi-agent paradigm is emerging for the development of solutions to very hard distributed computational problems. This paradigm is based either on the activity of "intelligent" agents which perform complex functionalities or on the exploitation of a large number of simple agents that can produce an overall intelligent behavior leading to the solution of alleged almost intractable problems.

Considering the complexity inherent to the manufacturing systems, dynamic scheduling is considered an excellent candidate for the application of agent-based technology. In many implementations of MAS systems for manufacturing scheduling, the agents model the resources of the system and the tasks scheduling are done in a distributed way by means of cooperation and coordination amongst agents [9-10]. When responding to disturbances, the distributed nature of multi-agent systems can also be a benefit to the rescheduling algorithm by involving only the agents directly affected, without disturbing the rest of the community which can continue with their work.

IV. NATURE INSPIRED OPTIMIZATION TECHNIQUES

Biological and natural processes have been a source of inspiration for computer science and information technology.

The interest of the Nature Inspired Optimization Techniques, also named Meta-heuristics, is that they converge, in general, to satisfactory solutions in an effective and efficient way (computing time and implementation effort). NIT have often been shown to be effective for difficult combinatorial optimization problems appearing in several industrial, economical, and scientific domains [3][9][11]. Prominent examples are evolutionary algorithms, simulated annealing, tabu search, scatter search, memetic algorithms, ant colony systems and particle swarm optimization.

When considering and understanding solutions followed by nature it is possible to use this acquired knowledge on the resolution of complex problems on different domains. From this knowledge application, on a creative way, arise new computing science areas - Bionic Computing.

The complexity of current computer systems has led the software engineering, distributed systems and management communities to look for inspiration in diverse fields, e.g. robotics, artificial intelligence or biology, to find new ways of designing and managing systems. Hybridization of different approaches seems to be a promising research field of computational intelligence focusing on the development of the next generation of intelligent systems.

V. AUTODYNAGENTS ARCHITECTURE

Distributed environment approaches are important in order to improve scheduling systems flexibility and capacity to react to unpredictable events. It is accepted that new generations of manufacturing facilities, with increasing specialization and integration, add more problematic challenges to scheduling systems. For that reason, issues like robustness, regeneration capacities and efficiency are currently critical elements in the design of manufacturing scheduling system and encouraged the development of new architectures and solutions, leveraging the MAS research results.

The work reported in this paper is concerned with the resolution of realistic scheduling problems, called here Extended Job-Shop Scheduling Problems (EJSSP) [9]. Moreover, it is concerned with integrated scheduling of jobs which are products composed by several parts or components that may be submitted to a number of manufacturing and multi-level assembly operations.

AUTODYNAGENTS is an Autonomic Scheduling System in which communities of agents model a real manufacturing system subject to perturbations. Agents must be able to learn and manage their internal behavior and their relationships with other autonomic agents, by cooperative negotiation in accordance with business policies defined by user manager.

It is a Multi-Agent system where each agent represents a resource (Machine Agents) in a Manufacturing System. Each Machine Agent must be able: to find an optimal or near optimal local solution through Genetic Algorithms, Tabu Search or other NIT; to deal with system dynamism (new jobs arriving, cancelled jobs, changing jobs attributes, etc.); to change/adapt the parameters of the basic algorithm according to the current situation; to switch from one Meta-Heuristic algorithm to another and to cooperate with other agents.

Scheduling approach followed by AUTODYNAGENTS system is rather different from the ones found in the literature; as we try to implement a system where each agent (Resource Agent) is responsible for optimizing the scheduling of operations for one machine through a NIT. This considers a specific kind of social interaction that is cooperative problem solving (CPS), where the group of agents work together to achieve a good solution for the problem.

The original Scheduling problem defined in [7][9], is decomposed into a series of Single Machine Scheduling Problems (SMSP)[9]. The Machine Agents (which has an NIT associated) obtain local solutions and later cooperate in order to overcome inter-agent constraints and achieve a global schedule.

Two possible approaches, to deal with this problem, could be used. In the first, the AUTODYNAGENTS system waits for the solutions obtained by the machine agents and then apply a repair mechanism to shift some operations in the generated schedules till a feasible solution is obtained (Repair Approach). In the second, a coordination mechanism is established between related agents in the process, in order to interact with each other to pursuit common objective through cooperation. These coordination mechanism must be prepared to accept agents subjected to dynamism (new jobs arriving, cancelled jobs, changing jobs attributes).

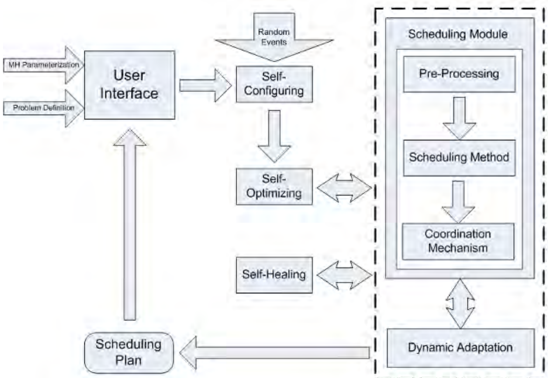


Fig. 1. AutoDynAgents Architecture

The AUTODYNAGENTS architecture (Figure 1) is based on six different types of agents. In order to allow a seamless communication with the user, a User Interface Agent is implemented. This agent, apart from being responsible for the user interface, will generate the necessary Task Agents dynamically according to the number of tasks that comprise the scheduling problem and assign each task to the respective Task Agent.

The Task Agent will process the necessary information about the job. That is to say that this agent will be responsible for the generation of the earliest and latest processing times, the verification of feasible schedules and identification of constraint conflicts on each job and the decision on which Machine Agent is responsible for solving a specific conflict.

The Machine Agent is responsible for the scheduling of the operations that require processing in the machine supervised by the agent. This agent will implement meta-heuristic and local search procedures in order to find best possible operation schedules and will communicate those solutions to the Task Agent for later feasibility check.

Respectively to the Self-* Agents, the Self-Configuring Agent is responsible for monitoring the system in order to detect changes occurred in the schedule, allowing the system to a dynamic adaptation. With this agent, the system will be prepared to automatically handle dynamism by adapting the solutions to external perturbations. While, on one hand, partial events only require a redefinition of job's attributes and re-evaluation of the objective function, on other hand, total events require changes on the solution's structure and size, carried out by insertion or deletion of operations, and also re-evaluation of the objective function. Therefore, under total events, the modification of the current solution is imperative, through job arrival integration mechanisms (when a new job arrives to be processed), job elimination mechanisms (when a job is cancelled) and regeneration mechanisms in order to ensure a dynamic adaptation of population/neighborhood.

The Self-Optimizing Agent is responsible for the automatically tuning of the meta-heuristics' parameters, according to the problem. This agent receives the initial problem, or the changes detected by Self-Configuring Agent, and automatically choose the meta-heuristic to use, and makes its self-parameterization. If some dynamism occurs, parameters may change in run-time. This tuning of parameters is made through learning and experience, since it uses a Case-based Reasoning (CBR) module. Each time a new problem (case) appears, the CBR uses past experience in order to specify the meta-heuristic and respective parameters for that case. When the new case is solved, it is stored for later use.

Finally, the Self-Healing Agent gives the capacity to the system for diagnosing deviations from normal conditions and proactively takes actions to normalize them and avoid service disruptions. This agent monitors other agents in order to provide overall self-healing capabilities. Since agents may crash for some reason, self-healing provides one or more agents backup registries in order to grant storage for the reactivation of lost or stuck scheduling agents with meaningful results, thus enabling the system to restart from a previous checkpoint as opposed to a complete reset. With this agent, the system becomes stable, even if some deadlocks or crashes occur.

Rescheduling is necessary due to two classes of events: Partial events imply variability in jobs/operations attributes such as processing times, due dates or release times; and Total events imply variability in neighborhood/population structure, resulting from new job arrivals, job cancellations, machines breakdown, etc.

VI. SELF-OPTIMIZINGMODULE

The objective is that each machine agent adopt and provides self-parameterization of the solving method in accordance with the problem being solved (parameters can change in run-time).

Each machine agent must be able to define which NIT will be used and define initial parameters of NIT according to the current situation or even to commutate from one algorithm to the other according to current state and previous learning. Agents self-optimize through learning and experience.

Nature Inspired Optimization Techniques like Tabu Search, Genetic Algorithms, Simulated Annealing, etc., are very useful to obtain good solutions in feasible execution times. Sometimes they can even obtain the optimal solution. But to be possible to obtain optimal or near-optimal solutions it is required the correctly tuning of the parameters from the different meta-heuristics. It was proved to be a very hard task since it needs some expertise knowledge from the meta-heuristic in use and from the concerned problem. Sometimes it needs the trial-error method and the difficulties increase when there are more than one meta-heuristic to be used, because it requires an a priori choice of the meta-heuristic and then the tuning of its parameters.

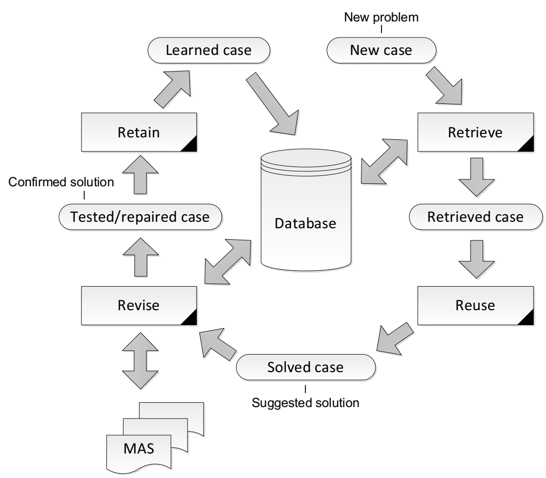


Fig. 2. CBR system architecture.

To solve this problem, we present a module, described in this section, representing the third part from Autonomic Computing Self-CHOP, named Self-Optimizing. Kephart and Chess [4] defined it as the capacity of "monitor, experiment with, and tune their own parameters and learn to make appropriate choices about keeping functions or outsourcing them". With this definition we propose an embedded module capable of monitoring the system and tuning the parameters of the different meta-heuristics, with respect to each new single problem emerging on the system.

This module needs to know how to tune up the different parameters of every meta-heuristic. It is impossible to predict every single problem to treat, so the system must be capable of learning about its experience during lifetime, e.g., as humans do. To perform this learning mechanism, we propose the use of a CBR system.

The CBR is an artificial intelligence methodology which aims to solve new problems by using information about the solutions to previous similar problems [5]. It operates under the premise that similar problems require similar solutions. The CBR consists in a cycle (Figure 2), usually described as the '4 Rs' [1]:

1. Retrieve the most similar case or cases;
2. Reuse the information and knowledge retrieved to solve the problem;
3. Revise the proposed solution;
4. Retain the revised solution to be useful for future problem solving.

There are two approaches to use the CBR in our system. The first one consists in retrieving the most similar case or cases to the new problem, regardless the meta-heuristic to use. Therefore, it is retrieved the case(s) containing the meta-heuristic and its parameters to use. The second approach consists in choose the meta-heuristic a priori and then retrieve the most similar cases, containing the parameters to use for that meta-heuristic. The first approach is more appropriate to the problem in question since that is important for the system to decide which meta-heuristic to use and the respective parameters, because not all meta-heuristics are suitable to every types of problems. Consequently, it will be possible to know, for example, which meta-heuristics are more suitable to a particular type of problem.

A. CBR Architecture

The architecture of the proposed CBR system is presented in Figure 2. Initially, every new problem leads to a new case for the system. In the first phase of the '4REs' cycle, the previous most similar cases are retrieved from the database. As it is possible to see in Figure 3, the retrieved cases are compared with the new case and two things can happen. If there is a very low similarity between them, the retrieved case will consist in a set of pre-defined parameters, tuned from expertise knowledge. If not, the most similar case is retrieved.

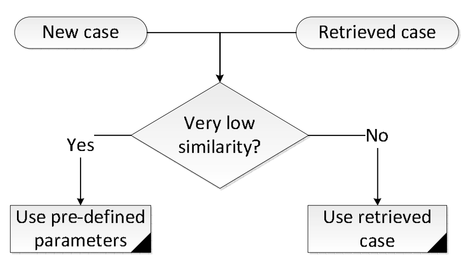


Fig. 3. CBR Retrieve phase

After being recovered, the case can be reused becoming in a suggested solution. After this, and in the Revise phase, the

CBR system executes the problem in the MAS, using the given solution. To escape from local optimal solutions and stagnation, we propose the use of some controlled randomness in the parameters of the proposed solution. After the MAS execution end, the case are confirmed or not as a good solution, and if so, it is retained on the database as a new learned case, for future use.

This represents the lifecycle of the proposed CBR system in the Self-Optimizing module, as each new problem will be evaluated and become a case of the casebase. We believe that the system will be capable of learning and evolving in the parameterization of the meta-heuristics, adapting to every new problem that may appear and improving the existing problems.

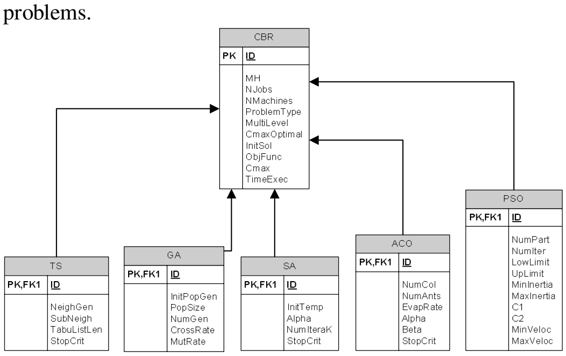


Fig. 4. CBR casebase

B. *Casebase*

In Figure 4 it is presented the CBR casebase. It consists in six tables, one for each meta-heuristic in use and one for saving the attributes of each case in the CBR.

Table I. CBR table fields description

Field	Data type	Description
ID	Integer	Primary key of the case, auto-incremented
MH	String	Name of the table of the meta-heuristic used
NJobs	Integer	Number of jobs of the problem (e.g., 10)
NMachines	Integer	Number of machines of the problem (e.g., 5)
ProblemType	String	The type of the problem (“single-machine”, “open-shop”, “flow-shop”, or “job-shop”)
MultiLevel	Boolean	Indicates if the problem has operations with more than one precedent
CmaxOptimal	Integer	Optimal makespan value known, if available
InitSol	String	Heuristic used to calculate the initial solution
ObjFunc	String	Objective function used
Cmax	Integer	Makespan value obtained from the case
TimeExec	Double	Execution time, in seconds

The NIT tables are TS, GA, SA, ACO, and PSO, representing Tabu Search, Genetic Algorithms, Simulated Annealing, Ant Colony Optimization, and Particle Swarm Optimization respectively. Each one of these tables contains the parameters of the cases for tuning the respective NIT.

The CBR table, as referred before, saves the attributes of every single case in the CBR system. The fields are described in Table I.

Table II – TS table fields description

Field	Data type	Description
NeighGen	Double	Neighborhood generation percentage
SubNeigh	Double	Sub-neighborhood percentage
TabuListLen	Integer	Tabu list length
StopCrit	Integer	Stopping criteria, number of iterations

Table III – GA table fields description

Field	Data type	Description
InitPopGen	Double	Initial population generation percentage
PopSize	Double	Population size percentage
NumGen	Integer	Number of generations
CrossRate	Double	Crossover rate
MutRate	Double	Mutation rate

Table IV – SA table fields description

Field	Data type	Description
InitTemp	Double	Initial temperature
Alpha	Double	Alpha factor of temperature reduction
NumIteraK	Integer	Number of iterations at the same temperature
StopCrit	Integer	Stopping criteria, number of iterations

As example we present in Table II, III, IV, V and VI the descriptions of the fields of each meta-heuristic, each field corresponding to each parameter tuned by the CBR system.

Table V – ACO table fields description

Field	Data type	Description
NumCol	Integer	Number of colonies
NumAnts	Integer	Number of ants per colony
EvapRate	Double	Pheromone evaporation rate
Alpha	Double	Heuristic value importance
Beta	Double	Pheromone importance
StopCrit	Integer	Stopping criteria, number of iterations

Table VI – PSO table fields description

Field	Data type	Description
NumPart	Integer	Number of particles
NumItera	Integer	Number of iterations
LowLimit	Integer	Lower limit
UpLimit	Integer	Upper limit
MinInertia	Double	Minimum inertia
MaxInertia	Double	Maximum inertia
C1	Double	Cognitive component
C2	Double	Social component
MinVeloc	Double	Minimum velocity
MaxVeloc	Double	Maximum velocity

C. Similarity Measure

The attributes to consider in the similarity measure are some fields from the CBR table. The most important ones are *NJobs*, *NMachines*, *ProblemType*, *MultiLevel*, and *CmaxOptimal*. These fields are correctly heavy in the similarity measure. However, there are other important attributes to consider for the selection of the most similar case to retrieve. These attributes are the Cmax value, because it is important to know the *makespan* of the previous cases, in order to be selected the best case from a set of similar cases, and the *TimeExec* value, which is important to know in those cases where exists very similar *makespan* values, in order to select the more efficient case. So, with these two values we can correctly heavy the effectiveness-efficiency binominal.

VII. CONCLUSIONS AND FUTURE WORK

In the retrieving phase, when there could be more than one similar cases, it will be retrieved the case with the best *makespan* value obtained in the shortest execution time.

It is expected a training phase of the CBR system, in which the system will be executed a number of times with some knowing problems only with the objective of case generation. Alternatively, there is the possibility to start without previous cases, using the pre-defined parameters.

We also predict that will be necessary to analyze and compare two different approaches of the CBR system utilization. The first approach, consists in apply the CBR system in a "global" level on MAS, optimizing the meta-heuristics in the same way for all Resource Agents, i.e., the CBR system specify the parameters before the start of MAS execution. The second approach is to consider the CBR in a "local" level of each Resource Agent, so each agent can optimize its parameters and/or apply a different meta-heuristic, according its single-machine problem. In this second case, every problem will be single-machine, though some Resource Agents can have different number of operations to process, which could be beneficial to consider. However, as a drawback, the *makespan* value to use will be the final time of processing in each machine, which can lead the agents to fall in local optimal, minimizing their own processing time, and not the global *makespan* of the MAS. First, we will implement the first approach, but with the intent to implement the second one, to be possible to compare and analyze each approach.

We believe that the resulting system will be able to evolve and improve the performance of the scheduling problems.

ACKNOWLEDGMENT

The authors would like to acknowledge FCT, FEDER, POCTI, POCI for their support to R&D Projects and GECAD - Knowledge Engineering and Decision Support Group Unit.

REFERENCES

[1] Aamodt, A. and Plaza, E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, Artificial Intelligence Communications, vol. 7, pp. 39-52, 1994.

[2] Aytug, H., Lawley, M. A, McKay, K., Mohan, S.& Uzsoy, R. Executing production schedules in the face of uncertainties: A review and some future directions. European Journal of Operational Research, Volume 16 (1), 86-110, 2005.

[3] Gonzalez, T. Handbook of Approximation Algorithms and Metaheuristics. Chapman&Hall/Crc Computer and Information Science Series, 2007.

[4] Kephart, J. and Chess, D. The Vision of Autonomic Computing, Computer, vol. 36, pp. 41-50, January 2003.

[5] Kolodner, J., Case-Based Reasoning, Morgan Kaufmann Publishers Inc, 1993.

[6] Luck, M., McBurney, P., Shehory, O., Willmoth, S. Agent Technology: Computing as Interaction. A Roadmap for Agent-Based Computing, AgentLink III, 2005.

[7] Madureira, A., Meta-Heuristics Application to Scheduling in Dynamic Environments of Discrete Manufacturing. PhD Dissertation. University of Minho, Portugal, 2003 (in portuguese).

[8] Madureira, A., Santos, F. e Pereira, I. Self-Managing Agents for Dynamic Scheduling in Manufacturing, GECCO'2008 (Genetic and Evolut. Comput. Conference 2008, Atlanta, Georgia (EUA), 2008.

[9] Madureira, A., Santos, J., Fernandes, N. and Ramos, C., Proposal of a Cooperation Mechanism for Team-Work Based Multi-Agent System in Dynamic Scheduling through Meta-Heuristics, 2007 IEEE Intern. Symp. on Assembly and Manufacturing (ISAM07), Ann Arbor (USA), 2007, pp. 233-238, ISBN: 1-4244-0563-7.

[10] Monostori, L., J. Váncza, Kumara, S. Agent based systems for manufacturing, CIRP Annals - Manufacturing Technology, vol. 55, no. 2, pp. 697–20, 2006.

[11] Xhafa, F. and Abraham A. Metaheuristics for scheduling in industrial and manufacturing applications series: Studies in computational intelligence, volume 128, Springer, 2008.

A Decomposition Method in Modeling Queuing Systems

Eimutis Valakevicius, Vaidotas Valiukas, Mindaugas Snipas
Department of Mathematical Research in Systems
Kaunas University of Technology
Kaunas, LT - 51368, Lithuania

Abstract-**The analysis of stochastic systems is notoriously hard, especially when their performance is described by multidimensional Markov chains. Approximation of general distribution functions by phase-type distributions rise to large Markov chains. If the model is too large to analyze in its entirety, it is divided into subsystems. Each subsystem is analyzed separately and global solution is constructed from the partial solutions. Numerical algorithm to solve for the steady-state probabilities of these Markov chains from a system of linear equations by decomposition method is presented in the paper. Based on these probabilities, we can compute a wide range of relevant performance characteristics, such as average number of customers of a certain type in the system and expected postponement time for each customer class.**

***Keywords*: Queuing system, Markov chain, decomposition method, steady state probabilities.**

I. Introduction

Complex queuing systems can arise in various practical applications, for example, telecommunication and computer systems, production, logistics etc. Queuing models are important tools for studying the performance of complex systems, but despite the substantial queuing theory literature, it is often necessary to use approximations in the case the system is non-markovian. Use of phase-type (PH) distributions is a common means of obtaining tractable queuing models [1-2], but it rises to large Markov chains. Markov modelling is a common approach to analyzing the performance of various stochastic systems. However, the performance analysis of complex queuing systems has a common source of difficulty: The Markov chain that models the system behaviour has a state space that grows infinitely in multiple dimensions. A method of decomposition to analyze Markov chains of large size is given in next sections.

It is known that creation of analytical models requires large efforts. Use of numerical methods permits to create models for a wider class of systems. The process of creating numerical models for systems described by Markov chains consists of the following stages: 1) definition of the state of a system; 2) creating equations describing Markov chain; 3) computation of stationary probabilities of Markov chain; 4) computation characteristics of the system performance. The most difficult stages are obtaining the set of all the possible states of a system and transition matrix between them. In the paper is used a method for automatic construction of numerical models for systems described by Markov chains with a countable space of states and continuous time.

To construct a model we need to describe the performance of a stochastic system in the event language [3]. It allows automating some stages of the model. The created software in C++ generates the set of possible states, the matrix of transitions among states, constructs and solves equilibrium equations to find steady state probabilities.

II. Approximation of Queuing System Models

Continuous-time Markovian processes with countable space of states are related with exponential distributions. It means that transition time from one state to another must be exponentially distributed. However, the mentioned condition often does not hold in the real life systems, so additional difficulties are expected to arise in the investigation and modeling of such systems. To solve this problem, an approximation of the stochastic systems by Markovian processes is used, when fictitious phase-type method is applied [4-6]. The diagram of approximation of a general distribution *G* by phase-type distribution is shown in Fig. 1.

The examples of approximation schemes of queuing systems are presented in figures 2 and 3.

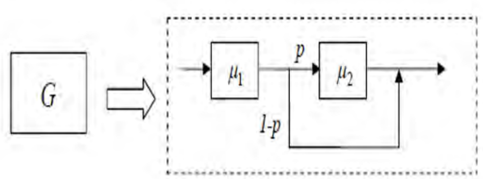


Fig.1. Approximation of general distribution G using phase-type distribution.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_75, © Springer Science+Business Media B.V. 2010

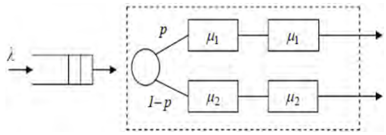


Fig. 2. Approximation of M/G/1 by mixture of Erlang distributions

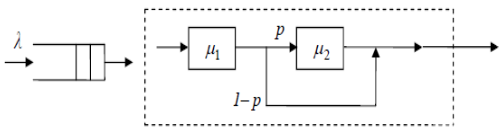


Fig. 3. Approximation of M/G/1 by Coxian distributions

III. CONCEPTUAL MODEL OF MULTICLASS QUEUING SYSTEM

Consider a multi-class, multi-server queuing system shared by N customer classes, numbered 1,…, N. A number of class indicates the priority rank (class 1 has highest priority and class N has the lowest priority). Priority rule is nonpreemptive. Class i customers arrives in the system according to a Poisson process with rate λ^i. The service times of class i customers are exponentially distributed with mean $1/\mu^i$. Within each class, service discipline of customers is a First Come First Served (FCFS). Queue length of each class customers can have various limitations.

IV. NUMERICAL MODEL OF THE SYSTEM

Consider a queuing system with 3 priority classes – high, medium and low, and 2 servers, which we call first and second. In addition, we consider limitation L ($L \in N$) for summary waiting space of each customer class. The system's performance is described in the event language, using methods from [1].

The set of system states is the following:

$N = \{(n_1, n_2, n_3, n_4, n_5)\}, n_1 + n_2 + n_3 \le L; n_4 \le 3; n_5 \le 3,$

$n_1, n_2, n_3, n_4, n_5 \in Z^+,$

where

n_1 – the number of high priority customers in the queue;

n_2 – the number of medium priority customers in the queue;

n_3 – the number of low priority customers in the queue;

n_4 – indicates the state of the first server (0 – if server is empty; 1 – if high priority customer is being served; 2 - if medium priority customer is being served; 3 – if low priority customer is being served);

n_5 – indicates the state of the second server (0 – if server is empty; 1 – if high priority customer is being served; 2 - if medium priority customer is being served; 3 – if low priority customer is being served);

The following events can occur in the system:

$E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9\},$

where

e_1 – a high priority customer arrived to the system with intensity λ^h;

e_2 – a medium priority customer arrived to the system with intensity λ^m;

e_3 – a low priority customer arrived to the system with intensity λ^l;

e_4 –a high priority customer was served in the first server with intensity μ^h;

e_5 –a medium priority customer was served in the first server with intensity μ^m;

e_6 – a low priority customer was served in the first server with intensity μ^l;

e_7 – a high priority customer was served in the second server with intensity μ^h;

e_8 – a medium priority customer was served in the second server with intensity μ^m;

e_9 – a low priority customer was served in the second server with intensity μ^l;

The set of transition rates between the states is the following:

$INTENS = \{(\lambda^h, \lambda^m, \lambda^l, \mu^h, \mu^m, \mu^l)\}$

The created software generates the possible set of states and matrix of transition probabilities.

Calculation of steady state probabilities can require a large amount of calculations and computer resources. For example, if the set of states is described as

$N = \{(n_1, n_2, n_3, n_4, n_5)\}, n_1 + n_2 + n_3 \le L; n_4 \le 3; n_5 \le 3,$

it is easy to prove that the total number of states equal to

$$|N| = \frac{(L+1)(L+2)(L+3)}{3!} \cdot 4 \cdot 4 = \frac{8}{3}(L+1)(L+2)(L+3).$$

For example, if $L = 10$ then system has 4576^2 states!

V. DECOMPOSITION OF THE MATRIX OF TRANSITION PROBABILITIES

Transition probability matrix P of Markov chain, which gives the conditional probability p_{ij} of making a transition from the state i to the state , j has the form

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{in} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nj} & \cdots & p_{nn} \end{pmatrix}$$

Let us assume that the matrix is of big dimension. So the calculation of the stationary probability vector π from

$$\pi \cdot P = \pi$$

becomes difficult, because classic methods and algorithms is not working properly due to the amount of computation and residuals. Therefore we will to apply the decomposition technique [9].

The algorithm of decomposition is presented in Fig.4.

In Markov modeling it is frequently the case that the state space of the model can be partitioned into disjoint subsets, with strong interactions among the states of the subset, but with weak interactions among the subsets themselves. Such problems are sometimes referred to as nearly completely decomposable (*NCD*) [7,8].

Strong interactions among the states of a group and weak interactions among the groups themselves imply that the states of a nearly completely decomposable Markov chain can be ordered so that the stochastic matrix of transition probabilities has the form

$$P = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1N} \\ P_{21} & P_{22} & \cdots & P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ P_{N1} & P_{N2} & \cdots & P_{NN} \end{pmatrix}$$

in which the nonzero elements of the off-diagonal blocks are small compared to those of the diagonal blocks. The sub blocks P_{ii} are square and of order n_i :

$$i = 1,2,\dots,N \,, \qquad n = \sum_{i=1}^{N} n_i$$

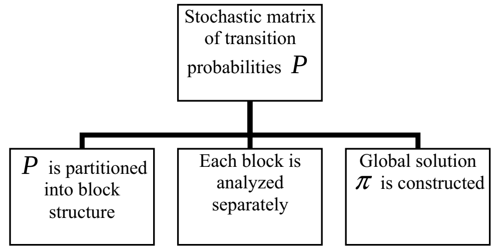


Fig.4. Scheme of the decomposition approach

The following three main steps are implemented during the partition of the matrix:

1. The decomposability factor γ is introduced, which varies from 10^{-10} to 10^{-1}.
2. All elements, less than or equal to γ are replaced by zero.
3. The resulting matrix is treated as the representation of a directed graph and a search is initiated for strongly connected components.

After all the result is the matrix *P* having the block structure.

Let π is partitioned respectively with *P*:

$$\pi = (\pi_1, \pi_2, \dots, \pi_N)$$

and π_i is a vector of length n_i. If the off-diagonal blocks are all zero, so the matrix P is completely decomposable and have the form

$$(\pi_1, \pi_2, \dots, \pi_N) \begin{pmatrix} P_{11} & 0 & \cdots & 0 & 0 \\ 0 & P_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & P_{N-1N-1} & 0 \\ 0 & 0 & \cdots & 0 & P_{NN} \end{pmatrix} = (\pi_1, \pi_2, \dots, \pi_N)$$

Each π_i can be found directly from

$$\pi_i \cdot P_{ii} = \pi_i \,.$$

In general case of nonzero off-diagonal blocks we assume that the system is completely decomposable and the stationary probabilities for each block are calculated in same way.

Another problem now arises. We have calculated the stationary probability vector from for each block. But simply connecting them together will not give a probability vector for all system. Why? The elements of each vector π_i sum to 1, while sum of the vector π this condition does not hold. We still need to weight each of the probability sub vectors by a quantity that is equal to the probability of being in that sub block of states. These weights are given by

$$\left(\|\pi_1\|,\|\pi_2\|,\ldots,\|\pi_N\|\right), \tag{7}$$

where $\|\cdot\|$ denotes the spectral form.

In purpose to compute the probability of leaving block i and entering block j the coupling matrix A is formed:

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{iN} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{Nj} & \cdots & a_{NN} \end{pmatrix}$$

where $a_{ij}=\dfrac{\pi_i}{\|\pi_i\|}P_{ij}e=\varphi_i P_{ij}e$, $e=(1,1,\ldots,1)^T$.

Thus, the element a_{1N} gives the probability that the system will enter one of the states of block N when it leaves one of the states of block 1.

Similarly in case of the matrix *P*, the stochastic matrix A possesses a unique stationary probability vector $\xi=(\xi_1,\xi_2,\ldots,\xi_N)$:

$$\xi\cdot A=\xi.$$

Only this time the element ξ_i is the stationary probability of being in block i (one of the states of block *i*). Furthermore, after some operations it becomes apparent that

$$\xi=\left(\|\pi_1\|,\|\pi_2\|,\ldots,\|\pi_N\|\right)$$

Now we have solutions π_i of each block and weights ξ_i of being in the block i from. So the global solution of the system can be constructed by

$$\pi=(\xi_1\pi_1,\xi_2\pi_2,\ldots,\xi_N\pi_N).$$

The noted steps are the basis of iterative *KMS* (Koury, McAllister and Stewart) and Takahashi algorithms [3]. They rapidly converge onto the exact solution when the Markov chain is nearly completely decomposable. Fig. 5 visually shows how the *KMS* algorithm is performed.

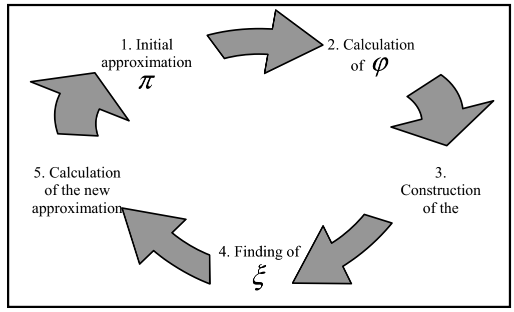


Fig. 5. The KMS algorithm

The Takahashi algorithm holds the same scheme of finding the solution as it is shown in the Fig. 5. The only difference between this algorithm and the *KMS* algorithm is in step 5, where the solution vector π is constructed (more on that in).

The software was created in order to realize this technique above. So let's look how decomposition approach works in G/M/1 and M/G/1 models. In purpose to model these systems let's say *G* is the Weibull distribution with the distribution function

$$p(t)=1-\exp\left\{-\frac{t}{\lambda}\right\}^k,\ t\geq 0,\ k>0,\ \lambda>0.$$

For $k=0{,}7$, $\lambda=0{,}7$ parameters of the mixture of Erlang distribution and Coxian distribution accordingly are:

$n=1$, $p=0{,}3106$, $\mu_1=0{,}5313$, $\mu_2=2{,}2869$;

$p=0{,}2385$, $\mu_1=2{,}2869$, $\mu_2=0{,}5313$.

In Fig. 6 we see the transition probability matrix *P* ($n=12$) of the model G/M/1 when the distribution function *G* is approximated by the mixture of Erlang distribution.

0	1	0	0	0	0	0	0	0	0	0	0
0	0	0,31061	0,68939	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
9,9999E-6	0	0	0	0	0,99999	0	0	0	0	0	0
0	9,9999E-6	0	0	0	0	0,310607	0,689383	0	0	0	0
0	0	0,653033	0	0	0	0	0	0,346967	0	0	0
0	0	0	0,304269	0	0	0	0	0,695731	0	0	0
0	0	0	0	9,9999E-6	0	0	0	0	0,99999	0	0
0	0	0	0	0	9,9999E-6	0	0	0	0	0,310607	0,689383
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0

Fig. 6. Transition probability matrix of G/M/1 (G – mixture of Erlang distribution)

There are only few elements less than 10^{-1} in the matrix (bolded in the picture), so the decomposability factor γ has no influence any more. If these probabilities (less than 10^{-1}) are replaced by zero as the theory expects, strong interactions don't exist in the model. Due to the similar structure of the transition probability matrix in the model G/M/1 (now *G* is approximated by the Coxian distribution), there are no strongly connected components (Fig. 7). Therefore, the first part of the decomposition method fails, because no block structure was found during this step. Some kind of the complement to the procedure of the partition is expected to be done.

During the analysis was noticed that both type of matrices have zero elements in the main diagonal. Moreover, this condition remains for sub matrices which were partitioned randomly. Elements p_{ii} in the transition probability matrix mean the possibility to stay at the same state. Since they all are equal to zero, every transition takes to another state.

Considering the remarks above, we can modify the procedure of the partition. Now we will be looking for sub matrices where all rows and columns have at least one nonzero element. The size of probabilities doesn't matter this time.

After applying the modification to G/M/1 model (*G* is Coxian distribution), we have got 3 sub matrices with the size of 4×4. All three blocks hold the condition about nonzero elements in every row and column. In G/M/1 model with the mixture of Erlang distribution the situation is different. The last sub matrix has only zero elements (Fig. 8), because it left as the residual of the procedure. In this case we can aggregate two last blocks into one. Now we have got 2 sub matrices with the size of 5×5 and 7×7, where at least one element in each row and column differs from zero.

0	1	0	0	0	0	0	0	0	0	0	0
0	0	0,238436	0,761564	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
9,9999E-6	0	0	0	0,99999	0	0	0	0	0	0	0
0	0,304269	0	0	0	0,165887	0,529844	0	0	0	0	0
0	0	0,653033	0	0	0	0,346967	0	0	0	0	0
0	0	0	9,9999E-6	0	0	0	0,99999	0	0	0	0
0	0	0	0	0,304269	0	0	0	0,165887	0,529844	0	0
0	0	0	0	0	0,653033	0	0	0	0,346967	0	0
0	0	0	0	0	0	9,9999E-6	0	0	0	0,99999	0
0	0	0	0	0	0	0	0,647165	0	0	0	0,352035
0	0	0	0	0	0	0	0	1	0	0	0

Fig. 7. Transition probability matrix of G/M/1 (G – Coxian distribution)

0	1	0	0	0	0	0	0	0	0	0	0
0	0	0,31061	0,68939	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
9,9999E-6	0	0	0	0	0,99999	0	0	0	0	0	0
0	9,9999E-6	0	0	0	0	0,310607	0,689383	0	0	0	0
0	0	0,653033	0	0	0	0	0	0,346967	0	0	0
0	0	0	0,304269	0	0	0	0	0,695731	0	0	0
0	0	0	0	9,9999E-6	0	0	0	0	0,99999	0	0
0	0	0	0	0	9,9999E-6	0	0	0	0	0,310607	0,689383
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0

Fig. 8. Block structure in G/M/1 model (G – mixture of Erlang distribution)

As was mentioned above, we didn't pay attention to the size of probabilities in sub matrices. But the theory requires that elements of the diagonal blocks are close to 1 as possible. Besides, the analysis of larger models showed some elements in the transition probability matrix repeat row by row (let's say in Fig. 7 elements $p_{4,5}$, $p_{5,6}$, $p_{6,7}$). These two conditions imply the fact that the partition in small matrices can be repeated in the large ones. So, the resulting strategy of the partition in the G/M/1 model is:

- P_{11} (size of 6×6), P_{22} (8×8), P_{33} (8×8),..., P_{NN} (smaller or equal to sub matrix of size 8×8), when G is the mixture of Erlang distribution.
- P_{11} (size of 5×5), P_{22} (6×6), P_{33} (6×6),..., P_{NN} (smaller or equal to sub matrix of size 6×6), when *G* is Coxian distribution.

These strategies of the partition are optimal due to they give faster convergence speed in comparison with previously mentioned ones. In Fig. 9 the example with the transition probability matrix of size $n=104$ shows that the number of iterations is less of the optimal strategy (line "Blocks2") than the first approach of the partition (line "Blocks1"). The significant difference between the accuracy wasn't noticed.

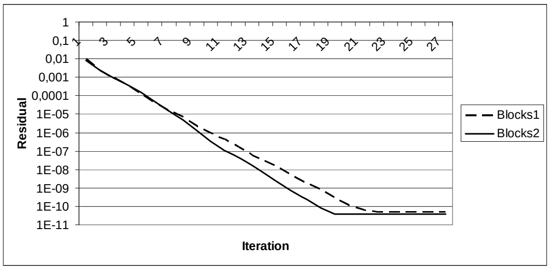


Fig. 9. Comparison of convergence in G/M/1 model

Applying the decomposition method with the optimal strategy of the partition into the block structure, the approximate stationary probabilities of large G/M/1 models were computed. The results require a lot space in the paper, so Table 1 shows average residuals. The accuracy (from 10^{-12} to 10^{-10}) remained as in models, where the size of the transition probability matrix was around 15. The number of iterations taken to reach this accuracy varies depending on the size of the matrix, but there were no cases when algorithms exceeded the number of 30 iterations.

VI. Summary

At the beginning of this paper we mentioned not only G/M/1, but also M/G/1 models (accordingly *G* is Coxian distribution or the mixture of Erlang distribution). But the analysis of these models showed that the different structure of the transition probability matrix doesn't let partitioning into sub matrices. After applying the procedure of the partition (even with simple condition – each row and column must have at least one nonzero element), resulting matrix is the same as primary transition probability matrix.

Queuing systems requires more investigation in the future.

TABLE 1.
Residuals of Large G/M/1 Models

Matrix size n	Average residual	
	KMS	Takahashi
123	1,4055E-12	1,6567E-12
243	3,5846E-011	9,6571E-012
483	6,2866E-011	7,9955E-011

References

[1] G. Mickevičius, E. Valakevičius , "Modelling of non-Markovian queuing systems", *Technological and economic development of economy*, vol . XII, No 4, pp. 295-300, 2006.

[2] T. Osogami , M. Harchol-Balter., "Necessary and Sufficient Conditions for Representing General Distributions by Coxians", *School of Computer Science, Carnegie Mellon University,* 2003.

[3] E. Valakevicius, H. Pranevicius, "An algorithm for creating Markovian Models of Complex System", *In Proceedings of the 12th World Multi-Conference on Systemics, Cybernetics and Informatics, Orlando, USA*, June-July 2008, pp. 258-262.

[4] M.A. Johnson, "An empirical study of queuing approximations based on phase-type distributions", *Commun. Statist.-Stochastic Models,* vol. 9, No. 4, pp. 531-561, 1993.

[5] J.A. Luhman, M.J. Johnson, "Behaviour of queuing approximations based on sample moments", *Applied Stochastic Models and Data analysis,* vol. 10, pp. 233-246, 1994.

[6] M. J. Faddy, "Examples of fitting structured phase-type distribution", *Applied Stochastic Models and Data Analysis,* vol. 10, pp. 247-255, 1994.

[7] S.W. Fuhrmann, R.B. Cooper, " Stochastic decompositions in the M/G/1 queue with generalized vacations", *Operations Research*, vol. 33(5), pp.1117-1129,1985.

[8] A. Horváth,, G. Horváth, and M. Telek., "A traffic based decomposition of two-class queueing networks with priority service", *Computer Networks*, vol. 53, pp.1235-1248, 2009.

[9] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*",.Princeton University Press, Princeton, New Jersey, 1994.

Modular Design and Structure for a Mobile Sensory Platform

Ayssam Elkady, Vipul Babariya, Jovin Joy and Tarek Sobh
School of Engineering
University of Bridgeport

***Abstract*—A mobile manipulator is a manipulator mounted on a mobile platform with no support from the ground. We are already in the process of building a platform (RISCbot II) which consists of a comprehensive sensor suite and significant end-effector capabilities for manipulation. In order to reduce the uncertainty in localization, sensor fusion is used to create an efficient and effective user interface to facilitate teleoperation by enhancing the quality of information that is provided to the teleoperator.**

This paper presents the modular design process of the RISCbot II mobile manipulator. In the design process, the overall design of the system is discussed and then the control process of the robot is presented. Furthermore, the tasks that the RISCbot II can perform such as teleoperation, navigation, obstacle avoidance, manipulation, face detection and recognition, and map building are described.

I. Introduction

A mobile manipulator offers the dual advantage of mobility offered by the platform, and dexterity offered by the manipulator. The mobile platform extends the workspace of the manipulator. We are developing and constructing a mobile manipulation platform called RISCbot II. The RISCbot II mobile manipulator platform is shown in figure 1. Mobile manipulators are potentially useful in dangerous and unpredictable environments such as construction sites, space, underwater, service environments, and in nuclear power stations.

Sensor fusion has been an active area of research in the field of computer vision and mobile robotics. Sensor fusion is the combination of sensory data from different sources, resulting in better and more reliable information of the environment than data derived from any individual sensor. Sensor fusion algorithms are useful in low-cost mobile robot applications where acceptable performance and reliability is desired, given a limited set of inexpensive sensors like ultrasonic and infrared sensors. Depending on the modalities of the sensors, sensor fusion can be categorized into two classes (as described in [1]): sensor fusion using complimentary sensors, and sensor fusion using competing sensors. Complementary sensors consist of sensors with different modalities, such as a combination of a laser sensor and a digital camera. In contrast, competing sensors are composed of sensors suit which have the same modality, such as two digital cameras which provide photographic images of the same building from two different viewpoints. Sensor fusion has some critical problems, including the synchronization of sensors. Different sensors have different resolutions and frame rates, so the sensors need to be synchronized before their results can be merged by fusing the data from multiple sensors, and presenting the result in a way that enables the tele-operator to perceive the current situation quickly. Sensor fusion is used to reduce the workload for the operator to enable him or her to concentrate on the task itself. Sensor fusion is commonly used to reduce uncertainty in localization, obstacle avoidance, and map building. Furthermore, sensor fusion may be used to improve teleoperation by creating user interfaces which efficiently facilitate understanding of remote environments and improve situational awareness.

Fig. 1. The RISCbot II mobile manipulator platform.

In this paper we describe our mobile manipulation platform, and introduce the tasks that the RISCbot II is performing them in, specifically teleoperation, navigation, obstacle avoidance, manipulation, face detection, and face recognition.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_76,

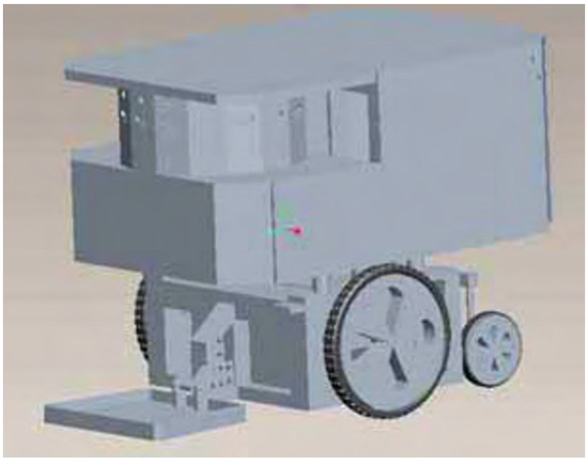


Fig. 2. The Prototype Structure.

II. DESIGN STRUCTURE OF THE RISCBOT II

In the final design as shown in figure 2, the three layer design was maintained. The base comprises of plywood holding a square core iron structure to provide support for the manipulator. Cylindrical structure to mount sensors is fixed on the central layer. The topmost layer is used to mount a stereo camera and a manipulator. The manipulator is mounted in such a way as to transfer its weight to the square core iron from the base.

Furthermore, there is a 2" diameter hole being drilled in the bottom layer to pass the wirings from the battery. The joystick of the wheelchair is mounted on the core iron structure vertically. There are seven sensors mounted on the back face of the bottom plywood and the center of all sensors is kept in one line. Different views of the structure are shown in figure 3. Plywood was chosen in the final design for building the layered structure due to its cost effectiveness and availability.

A. Individual Parts Description

A summary of individual parts used in the final prototype is described in table I.

TABLE I
INDIVIDUAL PARTS USED IN THE FINAL PROTOTYPE.

Part	Material Type	Quantity
Top Plywood Sheet	Plywood	1
Middle Plywood Sheet	Plywood	1
Bottom Plywood Sheet	Plywood	1
Central Support Structure	Core Iron	1
Support Part	Core Iron	4
Cylindrical Part	Wood	1
Backward Sensor Part	Wood	3

1) Bottom Plywood sheet: The bottom plywood sheet used has a dimension of 32" $\times$ 22" $\times$ 0.75" as shown in figure 4. In order to mount all components, thickness of the plywood sheet was set at 0.75". Front pocket is used to mount the laptop is

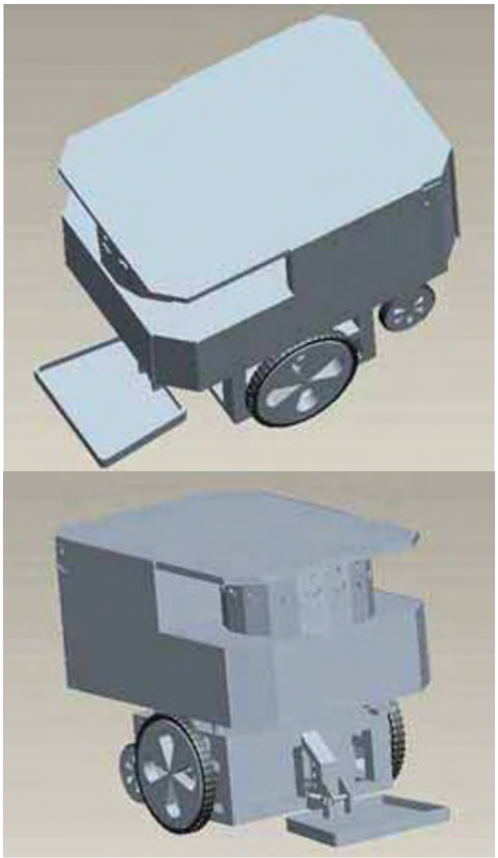


Fig. 3. Different views of the prototype structure.

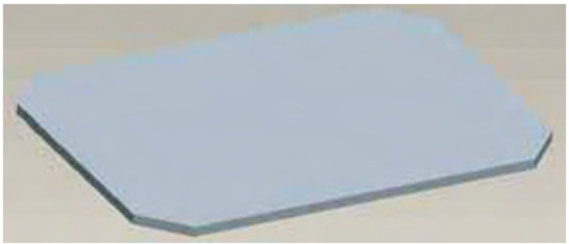


Fig. 4. Bottom plywood sheet.

shown in figure 5. This plywood sheet is directly mounted on six bosses extending from the wheelchair frame. Two I-support and Central support structure are mounted on top of the sheet.

2) Middle Plywood sheet: This plywood sheet shown in figure 6 is mounted on the square brackets of I-support and central support structure extending out of the bottom. Its dimensions are 16" $\times$ 22" $\times$ 0.75". This sheet is used to support the cylindrical sensor mount.

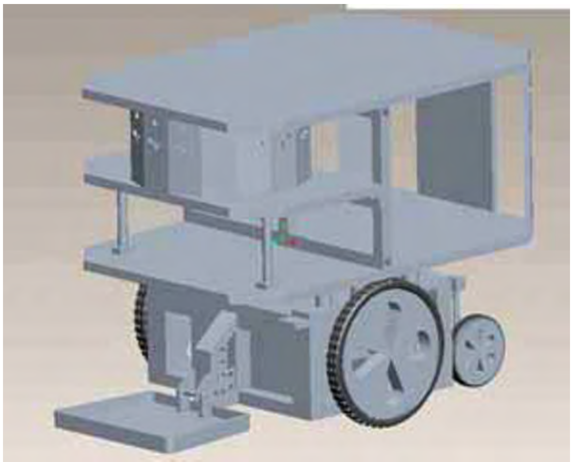


Fig. 5. The prototype showing structural support.

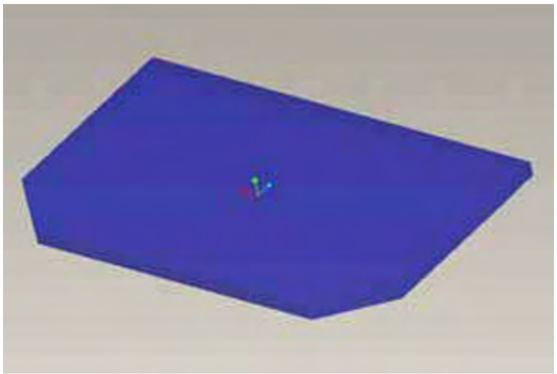


Fig. 6. Middle Plywood sheet.

3) I-Support Part: It is manufactured from core iron of 0.5" diameter with square brackets welded on each end as shown in figure 7. Four of these were used in the design; two of them are used for supporting the middle plywood sheet and the others are mounted on the middle plywood and is located in the Cylindrical Sensor mount to support the top plywood as shown in figure 8.

4) Cylindrical Sensor Mount: This structure is assembled from 9 rectangular wooden plates arranged around a incomplete circle (1.25π) of radius 9.5". Each rectangular plate is 2.75" thick and mounts 2 type of sensors, namely infrared and ultrasonic, as shown in figure 9.

5) Central Support Structure: This structure is a rectangular hollow frame made from core iron of 1" thickness. In the middle, an angle bracket is welded to support the middle plywood sheet. This support structure rests on the bottom plywood sheet and it provides support for the top plywood and manipulator.

Fig. 7. I-Support Part.

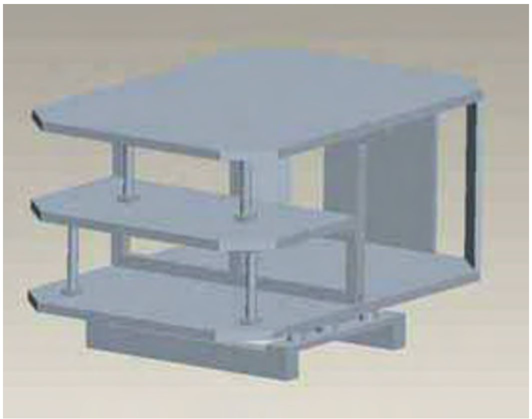


Fig. 8. The prototype showing structural support without Cylindrical Sensor Mount.

6) Top Plywood sheet: The top plywood sheet, shown in figure 11 is mounted on top of the central support structure. It is also supported by two I-support extended from the middle plywood, cylindrical sensor mount, and the rear sensor rack. This sheet supports the manipulator and stereo camera.

B. Analysis Of the final Structure

The analysis of the structure was done on ANSYS Workbench 11.0. Static structural analysis was done to check stress, strain, and deformation of the structure. The procedure used is described below:

1) Import the final prototype model designed using Pro-Engineer software into ANSYS.
2) Defining the material properties of each individual part

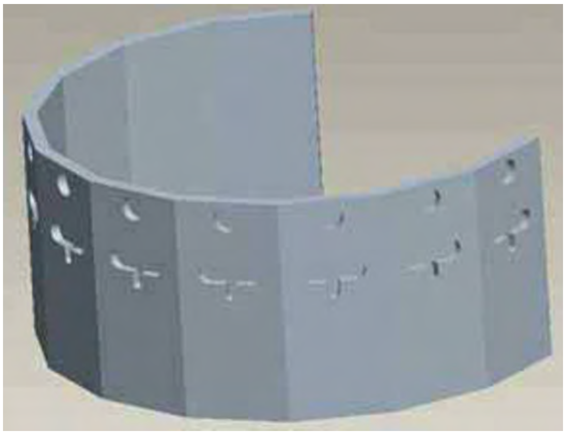


Fig. 9. Cylindrical Sensor Mount with rectangular plates on perimeter.

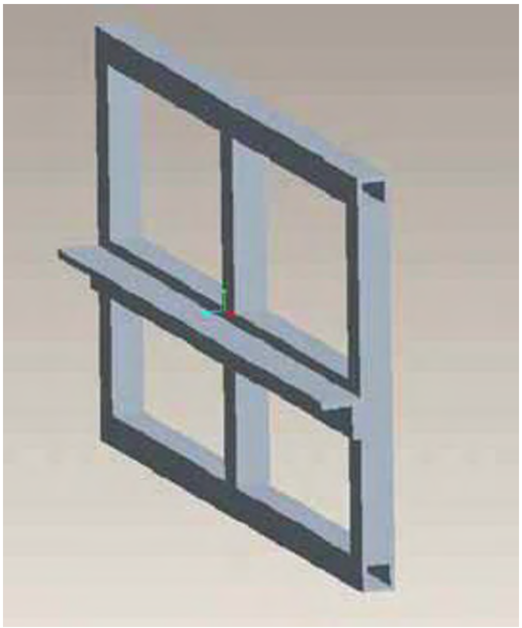


Fig. 10. Central Support Structure.

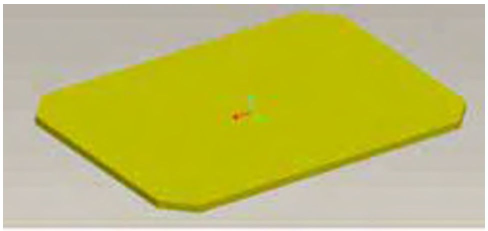


Fig. 11. Top Plywood sheet.

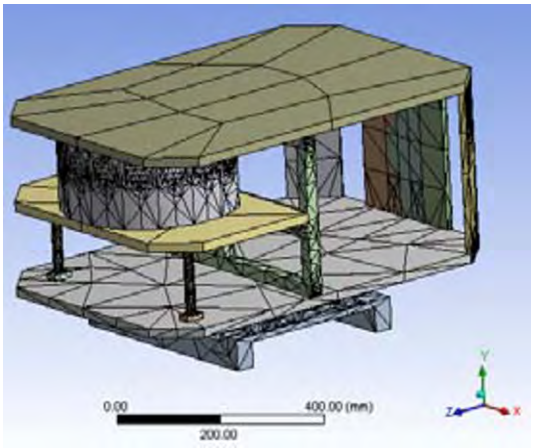


Fig. 12. Structure Meshing

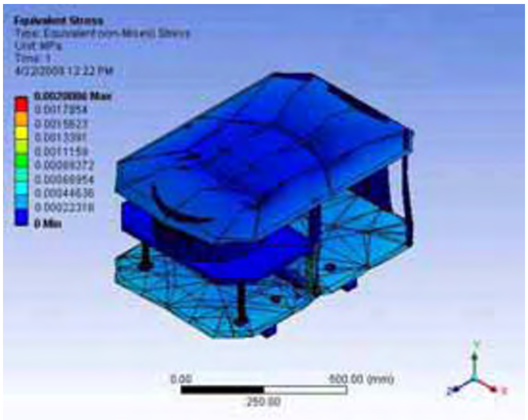


Fig. 13. Equivalent stress distribution throughout structure.

3) Assign meshes to the structure (figure 12)
4) Choose various support conditions and load application on the structure
5) Assign all the forces that affect the structure

The stress distribution in the structure is shown figure 13 and the maximum stress of 2.0086 kPa is measured on the edges of plywood. Strain generated in all the assembled components can be observed in figure 14. Total deformation found throughout the structure from the simulation was uniform as shown in figure 15. The maximum equivalent Elastic Strain and Total deformation found is negligible.

III. Applications

A. Face Detection

Efficiency of any detection process can be improved dramatically if the detection feature encodes some information about the class that is to be detected. Our primary method

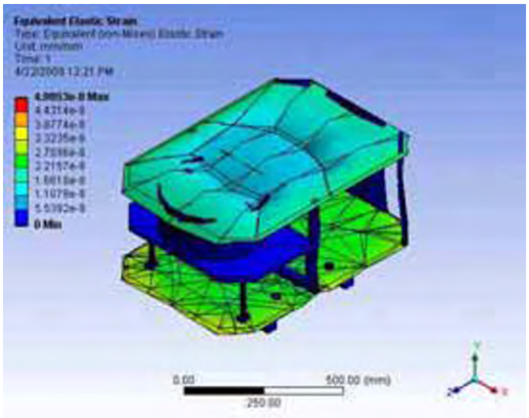


Fig. 14. Equivalent elastic Stain simulation result.

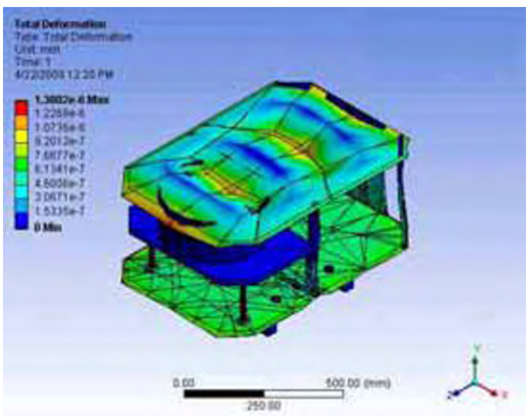


Fig. 15. Total deformation.

of detection is based on Haar-Link Features (as described in [2]) and OpenCV computer vision library. Haar encodes the existence of oriented contrasts between regions in images. A set of features can be used to encode the contrast exhibited by human faces and their special relationships.

Our detection program is derived primarily from OpenCV and uses classifiers (cascade of boosted classifiers working with haar-like features) that are trained with a few hundred sample views of human faces. Classifiers used in our project were obtained from OpenCV library. The classifier is built using positive and negative examples of the same size. After a classifier is trained it can be used to search across an image at locations. The output of the classifier is 1 if it finds an object (face) in a region. The classifier is resized, rather than the image, to detect objects (faces) of different sizes. Therefore, to find faces of unknown sizes, the scan procedure is done several times at different scales.

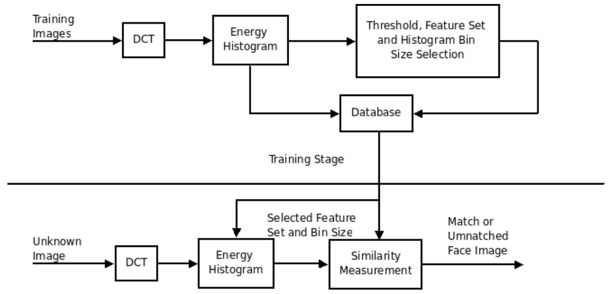


Fig. 16. Overview of Face Recognition Module

B. Face Recognition

Face Recognition is a task that identifies or verifies a person from a digital image or a video source. The primary method of face detection uses a overlapping energy histogram as described in Face Recognition by Overlapping Energy Histograms [3]. For the sake of efficiency, OpenCV library was used extensively while developing the code.

1) General Approach: Face Recognition algorithm is a two step process with a training stage and search stage. The general approach is shown in figure 16. The training stage is done initially to create a database which is not computed again as it is a time consuming process. The search phase uses the database created from the previous stage to look up unknown images. Overlapping energy Histogram comprises of taking DCT coefficients of every 8x8 pixel values with a 75 percent overlap as shown in figure 17 (i.e Taking DCT for every 8x8 block after every 2 pixels). A sample of which is shown in Figure 18. Once the image has been transformed into a histogram Euclidean Distances as show in Eqn 1 is computed, where Ω is the feature vectors calculated via the histogram. Further Euclidean Distance has been used in many face recognition techniques [3]. This distance is used to calcualte the Database Threshold(θ), which is used for identifying and rejecting images at the search stage

$$\epsilon_n = \|\Omega - \Omega_n\|^2 \tag{1}$$

C. Feature Extraction

Once the image with a face is transformed using overlapping DCT procedure values of Feature Set F2,

$$\mathrm{F2} = [\mathrm{DC}; \mathrm{AC}_{01}; \mathrm{AC}_{10}; \mathrm{AC}_{11}]$$

,is put into a histogram with a bin size of 60. The real challenge has been in selecting the appropriate appropriate bin sizes for calculating threshold values. A general deviation for the algorithm was using a predetermined bin size of 60 which proved to be optimally good, both computationally and for detection. This was determined by running a series of test on known images at the time of development.

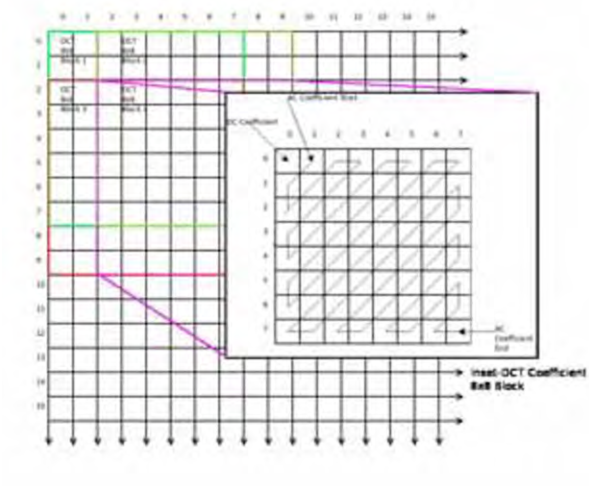


Fig. 17. Energy Histogram With 75 Percent Overlap

Fig. 18. Image and Corresponding Overlapping DCT Representation

1) Threshold Selection: Threshold for the image database is calculated by intra and inter class information gained from the training dataset. Intra class(D) information is obtained via computing the Euclidean distance between the images of an individual and inter class(P) information is obtained via computing the distance between images of an individual with others in the database. Once this is done for all the images in the database the Threshold(θ) is defined by

$$\theta = \frac{D_{max} + P_{min}}{2} \tag{2}$$

2) Image Search: Once the required feature from the image to be recognized is extracted, it's features are cross computed with each individual in the database to see if it falls withing the range of the database threshold. If more than one individual matches with the image that is searched for, the individual with the closest threshold value is usually the correct person.

Fig. 19. Face Detection Module Detecting Multiple People in a Single Frame

Fig. 20. Face Recognition Module Output

3) Face Detection/Recognition Testing: Face Detection and Recognition modules has to be tested simultaneously with the input of Detection Program fed into the recognition program. The Face Detection program on an average detects up to 5 faces real time(30 Frames/Second) running on a Dual Core Intel Processor, therefor bringing the total to 150 Images/Second. This is a 10 fold increase from the original Viola-Jones implementation. Figure 19 shows a sample output from the detection program. The main concern regarding the number of images that can be detected per frame is the computational requirement and the need to maintain real time performance.

Recognition module on the other hand can take each of the detected faces and search the database to find possible matches. The initialization process of the Face Recognition database was found to be a processor hog, hence plans to recompute the database values at run time had to be abandoned. Another bottleneck noticed was the total memory requirements for the database, which increased due to storing the feature vectors in uncompressed formats in system memory. Output of face recognition module can be seen in Figure 20.

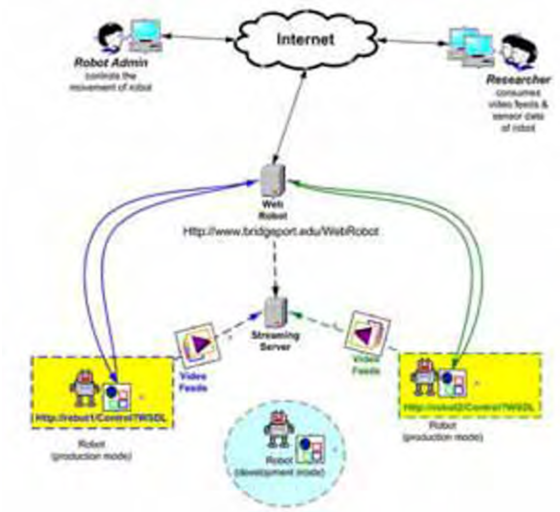


Fig. 21. RISCbot II website Architecture.

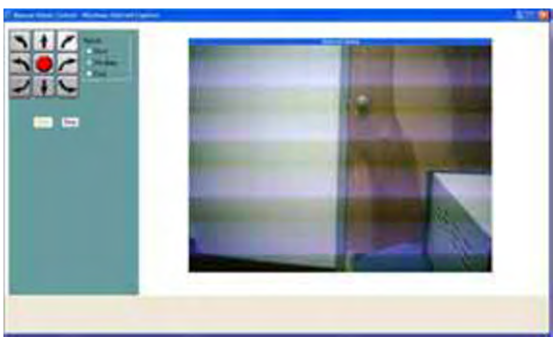


Fig. 22. Web-based manual control.

D. Teleoperation

We have implemented a web-based application in order to control the RISCbot II. The framework is implemented with ASP .NET 2.0 in $C\#$. The Web Form communicates with the physical robot using Web Services which are also implemented in $C\#$. In this application, a graphical user interface is provided to RISCbot's teleoperator to allow him to interact with the robot. There are two modes to control the robot; autonomous mode and manual mode. Since the RISCbot II could carry out a number of tasks, this web-based console allows the user to link the video/sensor data from the robot with various computer vision applications. The architecture of the web-based application for RISCbot II is shown in figure 21 and the manual control mode is shown in figure 22.

E. Manipulability

Studying the performance characteristics of the robot such as dexterity, manipulability, and accuracy is very important to the design and analysis of a robot manipulator. The manipulability is the ability to move in arbitrary directions. The manipulability index is considered as a quantitative and performance measure of the ability for realizing some tasks. This measure should be taken into consideration in the design phase of a serial robot and also in the design of control algorithms.

In [4], we presented a new method for measuring the manipulability index, and then justified this concept by visualizing the bands of this index resulting from our experiments implemented on different manipulators, such as the Puma 560 manipulator, a six DOF manipulator, and the Mitsubishi Movemaster manipulator. In fixed-base serial manipulators, manipulability depends on link lengths, joint types, joint motion limits and the structure of the manipulator. In mobile manipulators, the manipulability depends on the kinematics, geometric design, the payload, and the mass and mass distribution of the mobile platform. Thus, the manipulability measure in mobile manipulators is very complicated due to the coupling between the kinematic parameters and the dynamics effect. Furthermore, we use the proposed method for measuring the manipulability index in serial manipulators to generalize the standard definition of the manipulability index in the case of mobile manipulators.

F. Navigation and Obstacle Avoidance

A prerequisite task for an autonomous mobile robot is the ability to detect and avoid obstacles given real-time sensor readings. Given partial knowledge about the robot's environment and a goal position or a series of positions, navigation encompasses the ability of the robot to act based on its knowledge and sensor values so as to reach its goal positions as efficiently and reliably as possible. The techniques used in the detection of obstacles may vary according to the nature of the obstacle. The resulting robot motion is a function of both the robot's sensor readings and its goal position. The obstacle avoidance application focuses on changing the robot's trajectory as informed by sensors during robot motion. The obstacle avoidance algorithms that are commonly used can be summarized as the following [5]: the bug algorithm, tangent Bug, Artificial Potential Fields, and Vector Field Histogram.

G. Path planning and map building

Given a map and a goal location, path planning involves identifying a trajectory that will bring the robot from the initial location to the goal location. During execution, the robot must react to unforeseen obstacles and still reach its goal. In the navigation problem, the requirement is to know the positions of the mobile robot and a map of the environment (or an estimated map). The related problem is when both the position of the mobile robot and the map are not known. In this scenario, The robot starts in an unknown location in an unknown environment and proceeds to gradually build the map of the existing environment. In this case, the position of the robot and the map estimation are highly correlated. This problem is known as *Simultaneous Localization and*

Map Building (SLAM) ([6] and [7]). SLAM is the process of concurrently building a feature based map of the environment and using this map to get an estimation of the location of the mobile platform.

In [8], the recent Radio Frequency Identification (RFID) was used to improve the localization of mobile robots. This research studied the problem of localizing RFID tags with a mobile robot that is equipped with a pair of RFID antennas. Furthermore, a probabilistic measurement model for RFID readers was presented in order to accurately localize RFID tags in the environment.

IV. Implementation and Results

We are developing and constructing the mobile manipulator platform called RISCbot II(as shown in figure 1). The RISCbot II mobile manipulator has been designed to support our research in algorithms and control for an autonomous mobile manipulator. The objective is to build a hardware platform with redundant kinematic degrees of freedom, a comprehensive sensor suite, and significant end-effector capabilities for manipulation. The RISCbot II platform differs from any related robotic platforms because its mobile platform is a wheelchair base. Thus, the RISCbot II has the advantages of the wheelchair such as high payload, high speed motor package (the top speed of the wheelchair is 6 mph), Active-Trac and rear caster suspension for outstanding outdoor performance, and adjustable front anti-tips to meet terrain challenges.

In order to use the wheelchair as a mobile platform , a reverse engineering process has been used to understand the communication between the joystick of the wheelchair and the motor controller. This process was done by intercepting the continuous stream of voltages generated by the joystick after opening the joystick module and reading the signals within joystick wires that send the signals to the wheelchair controller.

We used different types of sensors so that the RISCbot II can perceive its environment with better accuracy. Our robot hosts an array of 13 $LV - MaxSonar^{®} - EZ0^{TM}$ ultrasonic sensors. The working envelope of the 13 sonars is shown in figure 23 The sensors are suitable for obstacle avoidance applications but their wide beams are unable to distinguish features within the beam angle, making sonars a poor choice of sensor for fine feature extraction within indoor environments. This resolution problem is magnified for objects further away from the robot (i.e., objects appearing at the wide end of the beam). Lastly, our robot is also equipped with an array of 11 Sharp GP20A21YK infrared proximity sensors above the sonar ring. The sonar and infrared sensors were mounted together so that their beams are oriented in the same direction. The configuration of sonar and infrared sensors is shown in figure 24. These sensors allow the RISCbot II to obtain a set of observations and provide these observations to the controller and higher decision making mechanisms. The controller acts upon the received observations to cause the robot to turn in the correct direction. The Integration of these modules together constitutes an intelligent mobile robot.

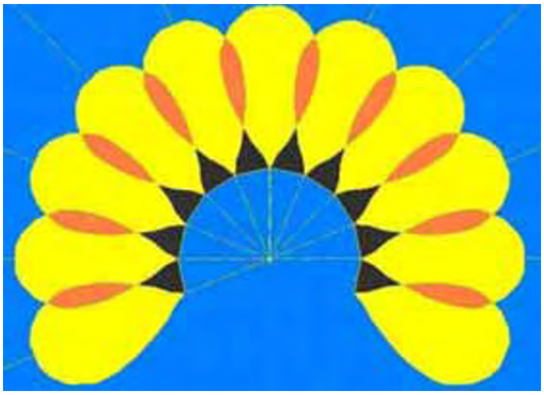


Fig. 23. Working Envelope of EZ0 sonar Sensors.

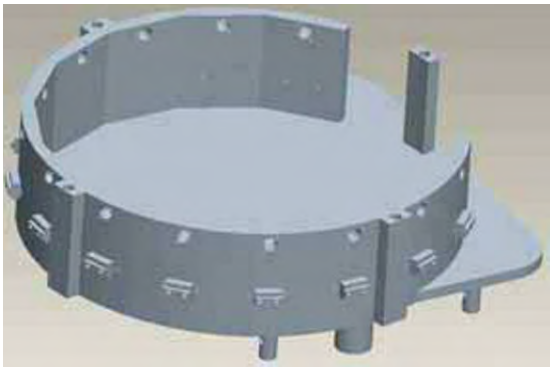


Fig. 24. A close-up view of the sonar and infrared sensors array.

A main drawback of these sensors is that they can only accurately measure obstacle distances within a range of 0.1m to 0.8 m. Another drawback of these sensors is that they are susceptible to inaccuracies due to outdoor light interference as well as an obstacle's color or reflectivity characteristics which can be seriously affected by windows and metallic surfaces.

Note that since our sonar and infrared sensors are in fixed positions, our experiments concentrated on performing data fusion on data obtained from a particular fixed height in the environment. In this project, sonar and infrared sensors are used together in a complementary fashion, where the advantages of one compensate for the disadvantages of the other.

As shown in figure 25, the RISCbot II software which is written in Visual $C\#$ and runs on a laptop reads the values of all sensors at a rate of 10 HZ gathered in the data acquisition. The RISCbot II software maps the sensory inputs to a series of actions which are used to achieve the required task. Based on the used algorithm, the RISCbot II software responds to the sensor data by generating streams of voltages corresponding to the joystick signals to the wheelchair controller. These voltages control the direction and the speed of the wheelchair causing

Fig. 25. The components of the RISCbot II system.

the RISCbot II to turn in the desired direction.

The experimental result indicates that the RISCbot II can detect unknown obstacles, and avoid collisions while simultaneously steering from the initial position towards the target position.

V. Conclusions and Future Work

In this paper, the mobile manipulation platform RISCbot II has been presented. The RISCbot II platform differs from any other robotic platform because its mo-bile platform is a wheelchair base. Thus, the RISCbot II has the advantages of the wheelchair. Furthermore, the RISCbot II consists of a comprehensive sensor suite, and significant end-effector capabilities for manipulation. In addition, we have used infrared and sonar sensors to monitor if any type of obstruction is in the path of the robot. This research aspires to find real-time collision-free trajectories for mobile manipulation platforms in an unknown static or dynamic environment containing some obstacles, between a start and a goal configuration.

Path planning for mobile robots is one of the key issues in robotics research that helps a mobile robot find a collision-free path from the beginning to the target position in the presence of obstacles. Furthermore, it deals with the uncertainties in sensor data.

The objective for this project is to implement a general purpose mobile manipulator that can be used for various applications such as teleoperation, navigation, obstacle avoidance, manipulation, 3-D reconstruction, map building, face detection, and face recognition. There are great benefits in using a mobile manipulator in dangerous, inaccessible and toxic environments. In teleoperation, a human operator controls the RISCbot II from a distance. The teleoperator has some type of display and control mechanisms, and the RISCbot II has sensors which gather all the information about the remote environment, an end-effector, and mobility.

In our anticipated future work, there will be an ongoing effort for the development of multiple mobile manipulation systems and platforms which interact with each other to perform more complex tasks exhibiting intelligent behaviors utilizing the proposed manipulability measure.

References

[1] A. Yilmaz, "Sensor fusion in computer vision," in *Urban Remote Sensing Joint Event*, April 2007, pp. 1–5.

[2] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2001*, vol. 1, 2001, pp. I–511–I–518.

[3] R. Tjahyadi, W. Liu, S. An, and S. Venkatesh, "Face recognition via the overlapping energy histogram," in *IJCAI*, 2007. [Online]. Available: http://www.ijcai.org/papers07/Papers/IJCAI07-465.pdf

[4] T. S. M. Mohammed, A. ElKady, "New concept in optimizing manipulability index of serial manipulators, using svd method," in *Industrial Electronics, Technology and Automation (IETA 07)*, Dec 2007.

[5] R. Siegwart and I. R. Nourbakhsh, *Introduction to Autonomous Mobile Robots*, 1st ed., ser. Intelligent Robotics and Autonomous Agents series. The MIT Press, April 2004.

[6] S. C. H. D.-W. M.W.M.G. Dissanayake, P.Newman and M. Csorba, "A solution to the simultaneous localization and map building (slam)problem," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229 – 241, Jun 2001.

[7] J. E. Guivant and E. M. Nebot, "Optimization of the simultaneous localization and map-building algorithm for real-time implementation," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 242 – 257, Jun 2001.

[8] D. F. K. F. Dirk Hähnel, Wolfram Burgard and M. Philipose, "Mapping and localization with rfid technology," in *IEEE International Conference on Robotics and Automation (ICRA 2004)*, vol. 1, New Orleans. LA, April 2004, pp. 1015 – 1020.

wradvs: A Didactic Server for IPv6 Stateless Autoconfiguration

Gabriel Abadi, Francisco Jimenez, and Eric Gamess
School of Computer Science
Central University of Venezuela
Caracas, Venezuela
gabriel.abadi@gmail.com, jimenezff@gmail.com, egamess@gmail.com

***Abstract* — In this paper we present *wradvs*, a free open source application developed under the GNU General Public License that provides a comprehensive solution for IPv6 stateless autoconfiguration. The application works as a service and implements RFCs 2460, 4861, 5006, 4191, 3775, and 3963. It also has an event log viewer that records detailed information of all events in real time, allowing users to troubleshoot network problems without the need of additional tools. The main goal of *wradvs* is to be used as a didactic application in network advanced courses at Central University of Venezuela. Thus, it has a friendly graphical user interface and presents a lot of valuable information to users while they configure all the parameters and options of the Router Advertisement messages.**

Keywords: ipv6; stateless autoconfiguration; router advertisement; windows; service; daemon; didactic application.

I. Introduction

Due to the imminent IPv4 address space depletion, LACNIC (Latin American and Caribbean Internet Addresses Registry) is encouraging the adoption of IPv6 [5][13] as soon as possible. However, this adoption has been slow in Venezuela and in many other countries. One of the reasons for the delay is the lack of IPv6 network specialists. Therefore, the training of IPv6 specialists has become an important issue. In the undergraduate program of Computer Science at *Central University of Venezuela* (in Spanish: Universidad Central de Venezuela), some courses have been upgraded or added to the curriculum to face the problem. For example, *Advanced Network Protocols* (in Spanish: Protocolos Avanzados de Redes) is a new course that was introduced to the curriculum of the undergraduate program of Computer Science in 2005. Its objectives include the understanding of IPv6 standards such as the stateless autoconfiguration process described in [1].

Some daemons or servers have been developed, by the community or manufacturers, to support the stateless autoconfiguration process. However, most of these daemons do not support all the parameters and options of the stateless autoconfiguration process, and sometimes, configuration can be very difficult and the debugging system can be very poor. These deficiencies can make the teaching and learning activity process very laborious. For these reasons, we developed *wradvs* (Windows Router Advertisement Server). *wradvs* is not only one of the most complete daemon created for IPv6 stateless autoconfiguration, but it also provides a friendly graphical user interface, assists users during the configuration process with tooltips, and validates all the submitted information. Additionally, *wradvs* provides an Event Log Viewer with detailed information of the messages that are sent and received in the autoconfiguration process. Even if the main goal of the application is to support the teaching and learning process, it can be very useful for a network administrator that uses IPv6 servers based on Windows. The rest of the paper is organized as follows: the IPv6 Router Advertisement message is presented in Section II; IPv6 stateless autoconfiguration is discussed in Section III; related works are viewed in Section IV; *wradvs* is presented and justified in Section V; and finally Section VI concludes the paper.

II. Router Advertisement

A Router Advertisement (RA) message is an Internet Control Message Protocol version 6 (ICMPv6) message defined by the Neighbor Discovery (ND) protocol [2].

IPv6 routers send unsolicited RA messages pseudo-periodically and solicited RA messages in response to the receipt of a Router Solicitation (RS) message. The interval between unsolicited advertisements is randomized to reduce synchronization issues when there are multiple advertising routers on a link. RA messages contain the information required by hosts to determine default gateways, the link prefixes, the link MTU, specific routes, home agent information, recursive DNS servers, whether or not to use stateful autoconfiguration, and the duration for which addresses created through stateless address autoconfiguration are valid and preferred [5].

Fig. 1 shows the structure of the RA message which contains the following fields:

- Type: the value of this field must be 134 and identify the type of ICMPv6 message (in this case, a RA).
- Code: the value of this field must be 0.
- Checksum: stores the checksum of the ICMPv6 message.
- Current Hop Limit: indicates the default value of the *Hop Limit* field to be put in the IPv6 header for packets sent by hosts that receive this RA message. A value of 0 indicates that the default value of the *Hop Limit* field is not specified by this router.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_77,

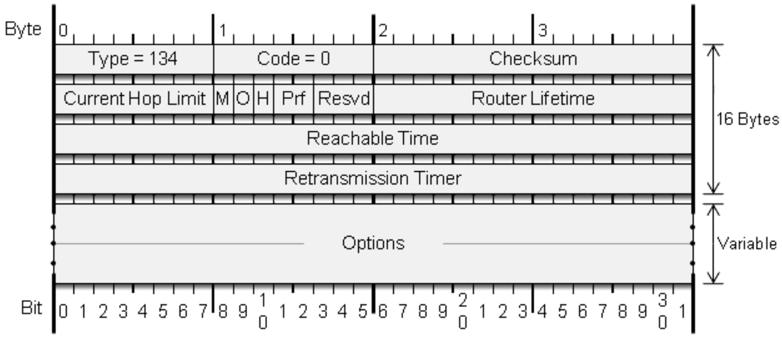


Figure 1. Structure of the Router Advertisement message.

- Managed Address Configuration Flag (M): when set, it indicates that hosts receiving this RA message must use a stateful configuration protocol, such as DHCPv6 (Dynamic Host Configuration Protocol for IPv6), to obtain addresses in addition to the addresses that might be generated from stateless address autoconfiguration.
- Other Stateful Configuration Flag (O): when set, it indicates that hosts receiving the RA message must use a stateful configuration protocol, such as DHCPv6, to obtain other configuration information.
- Home Agent Flag (H): when set, it indicates that the router sending the RA message is also functioning as a Mobile IPv6 home agent on this link [3].
- Default Router Preference (Prf): indicates whether to prefer this router over other default routers according to the level of preference. Valid values in binary are: 01 (high), 00 (medium) and 11 (low) [4].
- Reserved: this field is reserved for future use and must be set to 0.
- Router Lifetime: indicates the lifetime (in seconds) of this router as a default router. A value of 0 indicates that this router is not a default router; however, all other information contained in the RA message is still valid.
- Reachable Time: indicates the amount of time (in milliseconds) that a host should assume a neighbor is reachable after receiving a reachability confirmation. A value of 0 indicates that this router does not make any recommendation about the reachable time.
- Retransmission Timer: indicates the amount of time (in milliseconds) that a host should wait before the retransmission of a Neighbor Solicitation message. A value of 0 indicates that this router does not specify the retransmission timer.

The options that can be present in a RA message are the following:

- Source Link-Layer Address option: contains the link-layer address of the interface from which the RA message was sent.
- MTU option: contains the recommended MTU for the link. Should be sent on links that have a variable MTU.
- Prefix Information option: contains the prefixes that are on-link and/or used for stateless address autoconfiguration.
- Route Information option: contains more-specific routes that improve the ability of hosts to pick an appropriate next hop for an off-link destination [4].
- Recursive DNS Server option: contains one or more IPv6 addresses of recursive DNS servers [6].
- Home Agent Information option: contains information specific to this router's functionality as a home agent [3].
- Advertisement Interval option: contains the interval at which this router sends unsolicited multicast RA messages [3].

III. IPv6 Stateless Address Autoconfiguration

The IPv6 stateless address autoconfiguration defines the mechanism that allows a host to generate its own addresses and others network parameters using local information and information advertised by routers in RA messages. This mechanism requires no manual configuration of hosts. The address autoconfiguration process [1] for a physical interface of an IPv6 node is as follows:

When an interface becomes enabled, a tentative link-local address is generated based on the link-local prefix (FE80::/64) and an interface identifier, which is commonly derived from the link-layer address. The uniqueness of the tentative link-local address is verified using the duplicate address detection (DAD) algorithm. If the address is in use by another host, the autoconfiguration stops and manual configuration must be performed on the host.

Otherwise, the host sends a RS message to request an immediate RA message, rather than waiting for the next scheduled RA message.

The information advertised in the RA message (*Current Hop Limit*, *Reachable Time*, *Retransmission Timer*) is set. For each Prefix Information option sent in the RA message, the following actions occur:

1. If the On-Link flag is set to 1, the prefix is added to the prefix list.
2. If the Autonomous flag is set to 1, the prefix and an appropriate interface identifier are used to generate a tentative address.
3. Duplicate address detection is used to verify the uniqueness of the tentative address.
4. If the tentative address is in use, it is not assigned to the interface.

A RA message could also include other options, such as: Source Link-Layer Address, MTU, Route Information, Recursive DNS Servers, Home Agent Information, and Advertisement Interval.

IV. RELATED WORKS

A Router Advertisement daemon listens to RS messages and sends RA messages on demand and periodically as described in [2]. These advertisements allow any listening host to configure its addresses and some other parameters automatically through IPv6 stateless autoconfiguration.

We studied several Router Advertisement daemons of different operating systems to make a comparative study. The criteria for evaluating these daemons were:

- Installation: refers to the easiness of the installation process.
- Configuration: refers to the easiness of the configuration of parameters and options.
- Supported parameters: takes in consideration the amount of fully supported parameters and options.
- Graphical user interface: refers to the existence of a graphical user interface and its friendliness.
- Error handling: refers to how specific and clear the error messages are when there are troubles in the configuration.
- Event log: refers to the existence of event registry and how clear the events registered are.

The studied daemons were: *radvd*, r*tadvd*, *in.ndpd*, the Cisco IOS, and *netsh*.

Table I shows a summary of the study for all the Router Advertisement daemons. For each criterion, a grade was given between 0 and 5 stars, where 0 stars indicate a very poor or a lack of support, and 5 stars represent an optimum or efficient implementation.

radvd is one of the most popular Router Advertisement daemons. Most of the stateless autoconfiguration parameters and options can be specified with *radvd*. Configuration is done through a text file; there is no graphical user interface. The syntax of the configuration file is easy to read and learn. *radvd* provides an efficient and informative error handler; the depuration process shows information about the daemon's events. However, this information is not specific enough. The depuration information can be saved in a log file for posterior analysis. We tested *radvd* on Debian GNU/Linux 4.

rtadvd is another Router Advertisement daemon mostly used on FreeBSD and Mac OS. It supports most of the stateless autoconfiguration parameters and options. The syntax of the configuration file (text file) is hard to read and learn, the depuration information does not provides valuable information (e.g., value of the Router Advertisement's fields) and it does not allow a custom path for the log event. We tested *rtadvd* on FreeBSD 7.0 and Mac OS X Leopard.

in.ndpd was developed by SUN Microsystems for Solaris. It is compliant with RFC 2461 which was deprecated by RFC 4861. *in.ndpd* just supports an incomplete set of stateless autoconfiguration parameters and options. The debugging process displays important and well formatted information. It was tested on Solaris 10.

Some versions of the Cisco IOS (Internetwork Operating System) include a Router Advertisement daemon. This daemon supports just an incomplete set of stateless autoconfiguration parameters and options; however, it offers a very good debugging system. Since the IOS is a commercial product, no source code is available.

netsh is a command-line scripting utility that allows (among other things) sending RA messages in a Windows system. It does not have a graphical user interface. It only permits to specify a few parameters and options of the stateless autoconfiguration process. The error messages are very poor. The source code is not available.

TABLE I. COMPARATIVE ANALYSIS RESULTS.

	radvd	rtadvd	in.ndpd	Cisco IOS	netsh
Installation	★★★★★	★★★★★	★★★★★	★★★★★	★★★★★
Configuration	★★★☆☆	★★☆☆☆	★★☆☆☆	★★★★☆	★★★★☆
Supported parameters	★★★★★	★★★★★	★★★★☆	★★★☆☆	★★★☆☆
Graphical user interface	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆
Error handling	★★★★☆	★★★★☆	★★☆☆☆	★★★★☆	★★☆☆☆
Events log	★★★★☆	★★★☆☆	★★★☆☆	★★★☆☆	☆☆☆☆☆
Total	3.3	3.2	3.0	3.2	2.3

For teaching and learning purposes, support of all parameters and options of the stateless autoconfiguration process is an important issue. Another important issue is the availability of the source code so students can improve it and adapt it to specific needs. Additionally, a friendly graphical user interface can be very useful for non-expert students. Since none of the existing daemons offers all the required features, we decided to develop our own server.

V. WINDOWS ROUTER ADVERTISEMENT SERVER

wradvs (Windows Router Advertisement Server) is a Router Advertisement server for Windows developed under the GNU General Public License. It supports RFCs 2460, 4861, 5006, 4191, 3775, and 3963. It has a friendly graphical user interface, checks entered values for consistency, offers a good debugging system for troubleshooting purposes, and shows tooltips to guide users during the configuration process. At the present time, *wradvs* has an interface for English and Spanish users. Switching from one language to another can be done with just a few clicks.

A. Methodology

We wrote the application with Microsoft Visual C# Express Edition [8]. *wradvs* was divided in four main modules: the *wradvs Server*, the *Server Configuration Tool*, the *Log Viewer* and the *IPv6 Address Configuration Tool*. Fig. 2 shows a simplified class diagram of the application.

To develop *wradvs*, we used Scrum, an iterative incremental framework for managing complex work (such as new product development) commonly used with agile software development [7]. Scrum method is divided in Sprints. A Sprint is a period of time, where every phase of the development occurs. The four main modules were developed in the first four Sprints, starting with the *wradvs Server* that is installed as a service, and finishing with the construction of the detail of the events that occur in real time, showing a hexadecimal dump of the messages that are sent and received. Then, there was a Sprint for the creation of the Tooltips that gives all the information of the parameters, options and protocols that can be configured, making the application an excellent teaching tool. The following Sprint was introduced to improve the feedback to users and make the application more intuitive. Finally, there was a Sprint for testing and debugging all the software.

The application uses two external projects that were necessary to accomplish some functionality:

- WinPcap: an industry-standard tool for link-layer network access in Windows environments. It allows applications to capture and transmit network packets bypassing the protocol stack, and it has additional useful features; including kernel-level packet filtering and a network statistics engine [9].
- SharpPcap: a packet capture framework for the .NET environment based on the famous WinPcap libraries. It provides an API for capturing, injecting, analyzing and building packets using any .NET language such as C# and VB.NET [10].

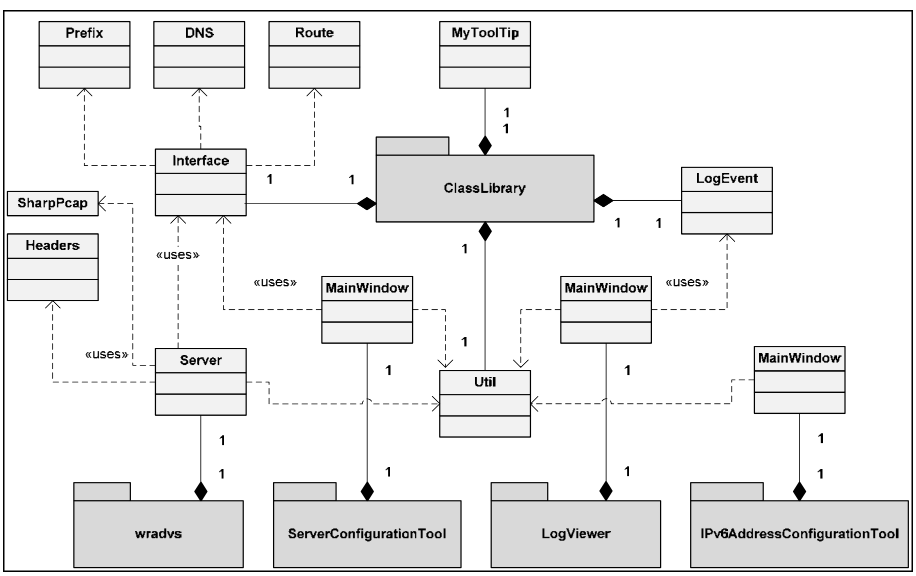


Figure 2. Diagram of the main classes.

B. Modules

As stated previously, *wradvs* is divided in four main modules (the *wradvs Server*, the *Server Configuration Tool*, the *Log Viewer* and the *IPv6 Address Configuration Tool*) that we describe below.

1) wradvs Server

The server (*wradvs*) is in charge of listening to the RS messages and sending RA messages on demand and periodically as described in [2], allowing any listening host on the link to configure its addresses and some other network parameters automatically through IPv6 stateless autoconfiguration. The server is installed on Windows as a service that will run on every system start.

2) Server Configuration Tool

The server configuration tool provides a friendly graphical user interface that allows users to configure all the parameters and options of the RA messages sent by the server. This configuration is saved as an XML document which is read by the server when it is started. The server configuration tool also provides a way to start, stop, and restart the server's execution. Since the application is intended for educational purposes, a lot of valuable information is shown on labels, tooltips, and messages. Also, this module checks all the data entered by users for consistency. Fig. 3 shows the main window of the server configuration tool.

3) Log Viewer

The log viewer provides visualization in real time of the server's events. For every event, the log viewer displays the following information: date, time, involved network interface, and description message. By double clicking on a specific event, the log viewer shows a hexadecimal dump (including every fields' values of the headers) of sent RA messages and received RS messages. Fig. 4 shows the main window of the log viewer.

4) IPv6 Address Configuration Tool

The IPv6 address configuration tool provides a graphical user interface that allows the execution of *netsh* commands related to the configuration of IPv6 addresses, default routers, routing table, DNS servers, and forwarding. Fig. 5 shows the main window of the IPv6 address configuration tool.

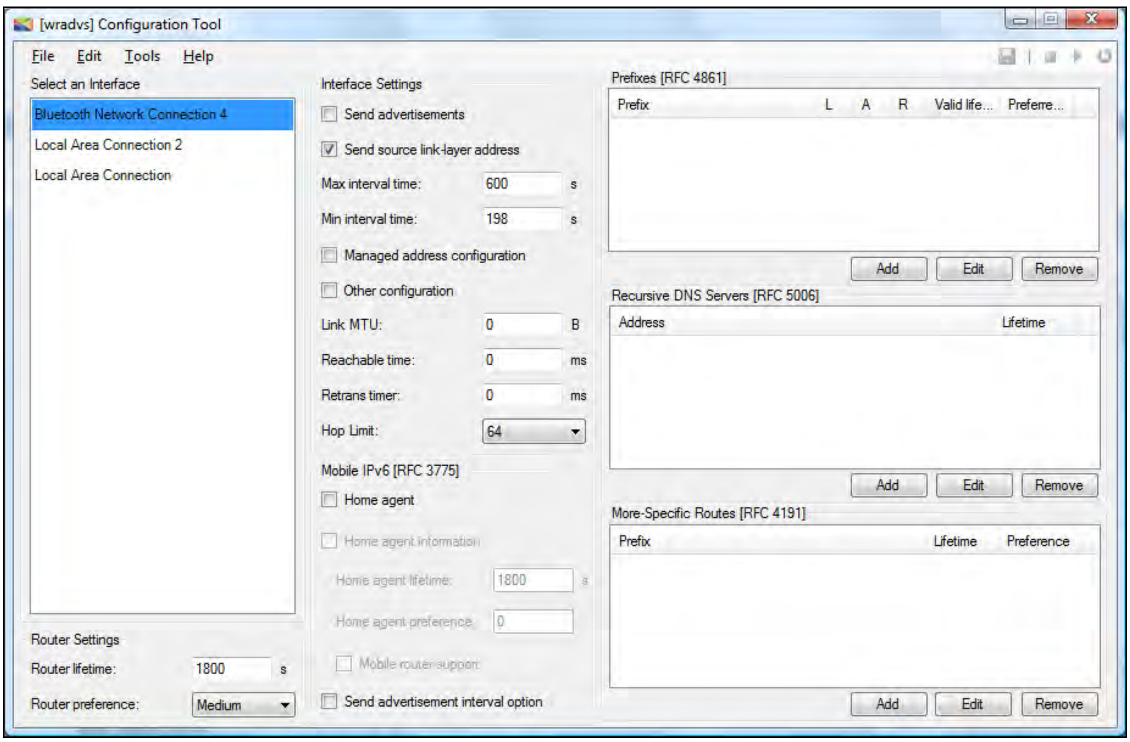


Figure 3. Main window of the server configuration tool.

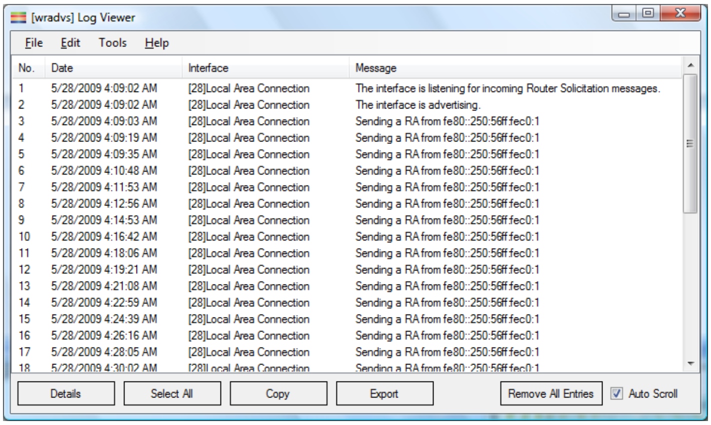


Figure 4. Main window of the log viewer.

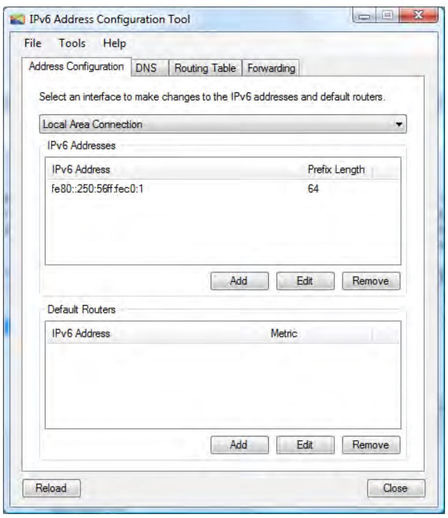


Figure 5. Main window of the IPv6 address configuration tool.

VI. CONCLUSION AND FUTURE WORKS

In this paper we presented *wradvs*, a full solution for the IPv6 stateless address autoconfiguration process. Thanks to the amount of information shown in the *Server Configuration Tool* and the amount of information stored in the *Log Viewer* for each event, *wradvs* is an excellent tool for teaching. Currently, it is used to teach the *Advanced Network Protocols* course at *Central University of Venezuela*. The feedback that we received from students and professors is very helpful and positive.

Although *wradvs* was primarily developed as a didactic application, it is a complete and powerful tool for network administrators. Its source code is freely available and can be easily adapted to specific requirements by students. It has been released recently and can be downloaded from Sourceforge at http://wradvs.sourceforge.net.

We will continue to support *wradvs*. Our goal is to offer it to other national and international universities to be used in their curriculum for teaching networking advanced courses. Thus, we plan to give support to many other languages (at the moment only English and Spanish are supported), and we also want to port *wradvs* to Linux. Since the stateless autoconfiguration process is vulnerable to various attacks [12] when it is not secured, the original specifications of ND called for the use of IPsec as a security mechanism to protect ND messages. However, the use of IPsec is impractical due to the very large number of manually configured security associations needed for protecting ND. For this reason, the Secure Neighbor Discovery Protocol [11] (SEND) was proposed. We also plan to develop an open source implementation of a SEND server that can be used in production environment or as a didactic application.

REFERENCES

[1] S. Thomson, T. Narten, and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862. September, 2007.

[2] T. Narten, E. Nordmark, W. Simpson, and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861. September, 2007.

[3] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6", RFC 3775. June, 2004.

[4] R. Draves and D. Thaler, "Default Router Preferences and More-Specific Routes", RFC 4191. November, 2005.

[5] J. Davies, "Understanding IPv6", 2nd ed. Microsoft Press. February, 2008.

[6] J. Jeong, S. Park, L. Beloeil, and S. Madanapalli, "IPv6 Router Advertisement Option for DNS Configuration", RFC 5006. September, 2007.

[7] K. Schwaber, "Agile Project Management with Scrum", Microsoft Press. March, 2004.

[8] P. Pelland. "Microsoft Visual C# 2008 Express Edition: Build a Program Now", Microsoft Press. March, 2008.

[9] http://www.winpcap.org

[10] http://sourceforge.net/projects/sharppcap

[11] J. Arkko, J. Kempf, B. Zill, and B. Nikander. "Secure Neighbor Discovery (SEND)". RFC 3971. March, 2005.

[12] P. Nikander, J. Kempf, and E. Nordmark. "IPv6 Neighbor Discovery (ND) Trust Models and Threats". RFC 3756. May, 2004.

[13] S. Deering and R. Hinden. "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460. December, 1998.

Securing P2P Wireless Communications by Deploying Honeytokens in a Cooperative Maritime Network

Jonathan White, Jamelle Brown
University of Arkansas
Fayetteville, AR USA
{jlw09, jab17}@uark.edu

Srini Ramaswamy, Sean Geoghan
Univ. of AR. at Little Rock
Little Rock, AR USA
{sxramaswamy, sjgeoghnan}@ualr.edu

Mhamed Itmi
INSA de Rouen
Rouen, France
itmi@insa-rouen.fr

***Abstract* - As wireless devices become ubiquitous in society, proper security methodologies must be deployed that are cognizant of the physical limitations inherent in the device design in order to ensure that the information produced in these systems is not abused. One such area of abuse involves an attacker locating and disabling a wireless device that is transmitting a relatively large amount of data compared to the other devices in the same physical area; this large amount of transmitted data may indicate that the wireless device is important or valuable to the owner.**

In this work, we propose and test a novel design that mitigates this threat as related to P2P wireless communication systems that have battery life as a limiting constraint. Our solution uses specialized honeytoken devices that transmit realistic noise in order to mask the presence or absence of communication from the actual wireless devices. This will allow the wireless devices to only transmit when required, conserving battery life and adding to the security of the system by making it impossible to locate a highly active device. The honeytoken noise is easily removed by trusted insiders; however it can not easily be removed by attackers. The idea of deploying honeytoken devices to protect P2P devices with battery life constraints is a new approach, and our key contribution is the proposal, design, and testing of this system.

I. INTRODUCTION

In the following sections we identify the maritime system in which we are deploying wireless peer to peer sensors in order to better track and secure intermodal shipping containers. We also define several key properties that all honeytokens possess, regardless of the type of object they are protecting. We also present the motivation for using honeytokens to better protect our wireless P2P system.

A. Maritime Shipping Industry

The shipping industry is a complex system, composed of many subsidiaries that allow and facilitate the transport of goods from manufacturers to retail outlets or other entities in the supply chain. In this work, we are focused on one particular subsystem, the maritime shipping industry, which involves the movement of goods from large centralized ports to other ports by utilizing cargo ships that traverse the oceans and other waterways. The maritime shipping industry relies on shipping freight that is containerized, meaning that the cargo is sealed by the owner in a standardized metal container and is not palletized or packaged by the maritime shipping provider. This containerization has allowed the shipping industry to achieve great efficiencies as the whole system has become standardized.

There are only a few standard sizes of containers that are used throughout the industry (e.g., 8'x8'x20', 8'x8'x40', etc). The container shipping industry measures freight in terms twenty foot equivalent units (TEUs), which represents the capacity of a single 20'x8'x8'6" container. Estimates for 2008 indicate that over 18 million containers make over 200 million trips per year, which account for approximately 90% of non-bulk cargo shipped worldwide [6]. Currently, the largest container transport ship is the Emma Maersk which can transport over 11,000 20-foot containers [5]. The ports are impressive locations, for example, the Port of Los Angeles encompasses 7500 acres, 43 miles of waterfront and features 27 cargo terminals.

Container shipping via large cargo ships is an industry that transports millions of tons of cargo on hundreds of cargo ships to destinations all over the world. In a typical journey, a shipper loads goods into a container and then seals the container. Next, the container is transported via truck to a port where it is loaded onto a container ship for overseas transport. Once the ship arrives at its destination port, the container is unloaded and again loaded onto a truck for transport to its final destination where it is unsealed and unloaded. Due to the value of the cargo and the potential for terroristic actions such as smuggling contraband, poisoning food cargo or stealing contents, the container must remain secure throughout its entire journey. Also, the container must remain at a safe temperature for refrigerated goods and the containers should not experience a high level of acceleration for fragile goods. A system that can monitor the position of the container at a port or on a ship

K. Elleithy et al. (eds.), *Technological Developments in Networking*, *Education and Automation*,
DOI 10.1007/978-90-481-9151-2_78, © Springer Science+Business Media B.V. 2010

and whether or not the container has been breached will provide shipping personnel with additional information to ensure the secure transport of the containers and the goods within the container. Our system provides part of this security and is described in section 3.

B. Honeytokens

Honeytokens are an information security tool that operates on the theory of deception. Honeytokens are related to honeypots, which are systems whose value lies in attracting, distracting, and potentially tracing attackers by presenting a seemingly worthwhile, but unprotected, computer. Honeytokens are different from honeypots in that they aren't an actual device. Instead, they are seemingly worthwhile and valid pieces of digital information. While organizations have sophisticated firewalls and intrusion detection systems in place to detect physical attacks on their networks or data systems, the loss of information posses a much greater threat, especially in wireless systems like ours [1].

The term 'honeytoken' was coined by Augusto Paes de Barros in 2003 in a discussion on honeypots. Honeytokens are manufactured by security engineers to look attractive to an attacker [1], hence the name. Examples would be a customer database of personal information, a list of passwords, proprietary design documents, a list of credit card numbers, or a seemingly worthwhile wireless signal passing through a P2P network.

Honeytokens have common properties, regardless of the digital object they mimic. First, the honeytoken must appear valuable to an adversary [2]. If the honeytoken does not appear worthwhile, it will never be interacted with, stolen, or examined. As wireless systems become ubiquitous, the information that they produce will become valuable as it will be used in many different areas to communicate important data.

Second, the honeytoken must be designed in such a manner so as to not interfere with any real data. The implementers of the honeytokens must be able to tell what is real and what is not, and thus the honeytoken must have unique, identifiable properties. If the honeytokens can not be separated from the real data, the whole system suffers and the accuracy of the system is compromised.

Third, honeytokens are implemented so that they appear real to an attacker. If the honeytokens are obviously an imitation, the attacker would have no reason to use or examine them. If the honeytoken is easily detected, it might also be easily removed by the attacker while still accessing the actual data that the honeytoken was designed to protect. Secrecy is of the utmost importance because if an attacker is able to find out what objects are honeytokens, they can easily ignore them [3].

C. Motivation

The system in which we are operating consists of many distributed sensors that communicate in a peer to peer manner. The wireless system encrypts all communications, making it unintelligible to attackers. Also, due to the inherent battery constraints, the devices will only transmit when they have information to communicate or are passing on information received from another node. This allows for fault tolerance and long battery life, which are two requirements of the system.

However, the proposed system exposes a security risk. While an attacker may not be able to understand what the devices are communicating, they can easily locate the devices that are communicating the most as most devices will be in a sleep state the majority of the time. For example, when a container is breached it will raise a signal that will propagate through the system. As this is a rare event, if an attacker has been monitoring the wireless communications with a high powered receiver and suddenly notices a spike in the received traffic, they can be assured that something important has happened. Also, if a device is consistently being queried for information by the owner, this device will also communicate relatively more data than the other containers in the P2P network, making it stand out from the rest. An intelligent and motivated attacker could potentially target this container for malicious actions, such as stealing the contents, as it appears to be valuable to its owners.

We are going to use honeytokens to mitigate this risk. The honeytokens will be deployed throughout the ship and they will produce realistic looking noise; the communication from the honeytokens will be very similar to the actual communications produced by the sensors. Each honeytoken device will be given a unique identification value so that they can be separated from the actual data produced by the P2P sensors. The honeytoken identification values will be known by the controlling system, and the controlling system will notify the sensors to not pass the noise produced by the honeytokens through a learning phase when the ship is loaded. The noise will appear at random times and at random durations, effectively masking the actual signal communications that the sensor nodes produce. Because there will always be a high level of realistic noise in the RF environment, an attacker will no longer have the ability to single out an individual container, adding to the security of the system. The use of honeytokens also allows the actual P2P sensors to transmit only when they have actual information to send, conserving battery life.

The rest of this paper is organized as follows. Section 2 details some previous work in the area of P2P security and honeytokens. Section 3 describes the current P2P wireless system and our proposal for using honeytokens. Section 4 shows how we simulated the system and took comparison measurements on a system that uses honeytokens and with a system that does not, and section 5 presents these results. Section 6 concludes the work.

II. BACKGROUND

Honeytokens are used in several diverse industries [4]. They have been used with great success, and this is an indication that P2P honeytokens can potentially be used in the maritime shipping domain. We show what industries use honeytokens and also other honeypot methods that have been used to secure P2P systems.

Honeytokens have been used in personally identifiable information (PII) databases to detect unauthorized access [1]. In many industries, privacy laws prohibit individuals from accessing personal information that they don't have a valid use for. Detecting when unauthorized access has occurred is difficult, as the data requirements for applications and users vary greatly.

Honeytokens are known as 'copyright traps' when used in the cartographic and dictionary industries [3]. Mapmakers are known to insert fake streets, misspelled objects, and non-existent bends in rivers to catch illicit copying. Dictionaries have also inserted fake words into their dictionaries in order to catch copyright violators. Both of these uses of honeytokens have been shown to work [3].

Honeypot systems have been used to analyze P2P botnets [7]. The honeypot attracts the bots and the actions that are performed are analyzed. If the bot interacts with the honeypot for an extensive period of time, many of its abilities can be identified and analyzed. Honeytoken email addresses have been used to monitor distributed P2P spam networks [8]. The addresses appear valid, but they are present just to attract and waste attacker's time and resources. Many other approaches have been used to secure P2P communications [9], [11], but little work has been done on using honeytokens to add noise to the system and to distract attackers from the actual communications.

III. PROPOSED SYSTEM

In the following sections, we describe the P2P wireless maritime system. As the sensors are going to be deployed onto shipping containers, there are several physical constraints that must be accounted for. Also, we describe in detail our proposal for deploying the honeytoken devices throughout the RF environment on the ship.

A. P2P Wireless Maritime System

The container network is created by instrumenting each container with intrusion detection hardware and a wireless network device. The device will have a battery, radio, temperature sensor, light sensor and accelerometer in order to perform the necessary physical security requirements. Due to the limited range of the wireless devices, many containers will not be able to communicate directly to the ship or port network. However, the containers are packed tightly together and each container will be in range of several neighbor containers. Typically, containers are stacked several units high in rows on the ship or port.

Additionally, some of the containers will be in range of the ship or port network. Therefore, all co-located containers, either on a ship or a port, will form an ad-hoc peer-to-peer (P2P) network. Part of the cost-effectiveness of the containerization shipping process is the fact that individual containers are designed to be rugged and require very little maintenance. Therefore, the addition of sensitive electronic equipment that requires frequent maintenance will not perform well within the economic constraints of the shipping industry. Solid state electronic components are generally rugged enough to withstand the application as long as they are mounted to the container in a position that is not vulnerable to damage by the equipment used for loading and unloading. However, the power requirements of the instrumentation must be addressed. Typically, containers do not have access to a direct, always on, power source. Therefore, each wireless node must rely on battery power, which must last several months to prevent excessive maintenance requirements.

Each container node periodically broadcasts its status (breached or secure) to all the container nodes that are listening. For example, the sensor nodes may broadcast a system secure message every hour on the hour; also the sensors will wake every ten minutes to listen for breaches or other messages from the system. If a breach message is received, the message is passed on throughout the P2P network until it reaches the intended recipient. This is a basic overview of the P2P environment in which we are operating in. While the focus of this work is not the implementation of the network itself, it is necessary to have an appreciation of the physical constraints on the devices in order to understand the security risks inherent in this design.

B. Honeytoken Deployment

The honeytokens will be deployed throughout the RF environment on the boat. They must be deployed in sufficient amounts in order to effectively cover the actual communications that the sensors are transmitting. Also, as the sensors are synchronized to transmit and/or receive only at certain predetermined discrete intervals, the honeytokens must also be attuned to this. The honeytokens must produce noise that appears realistic to an outsider when compared to the actual transmissions, however since the communications are all encrypted, this will be a straightforward process.

The honeytoken devices themselves will be simpler than the P2P devices. The honeytokens will have a transmission only radio, a battery, and a small processor that can store the device's assigned identification number and perform the necessary calculations in order to produce random noise at a specified time slot. The honeytokens will run on a separate battery based power supply from the P2P devices so as to not have a negative effect on the battery of the actual devices. There will be no sensors or other physical interfaces as the honeytokens are present to only produce noise.

The honeytoken devices can be placed on physical containers or throughout other areas of the ship. When the honeytokens are placed throughout the ship, the controlling computer records the identification number of the honeytoken and notifies collocated devices about their presence. If an actual P2P device retransmits the noise, it will eventually reach the ship's computer. The computer will have access to a list of the honeytoken identification values and it will be able to determine that this message is in fact noise. The shipboard computer will then send a message to the device that retransmitted the noise alerting it to the fact that this device is in fact a honeytoken. In this way, all trusted devices can learn what devices are actual sensors and which devices are present just to produce noise. This method of learning what devices are honeytokens will require extra communication during the setup phase, but once the honeytokens are learned, the P2P devices can ignore the noise while still being able to retransmit valid messages.

As mentioned previously, honeytokens must appear valuable, realistic, and must not interfere with the valid data. The honeytokens on our system are valuable and realistic because they produce the same type of signals that the actual P2P devices transmit. An attacker does not have the ability to remove this noise, and they can no longer locate a signal device that is transmitting a large amount of data. As each of the honeytokens has a unique identification value that is only known to the trusted devices, each device can correctly identify what is and what is not noise, keeping the devices from interfering with each other.

IV. SYSTEM SIMULATION

In the following section, we detail the steps that were taken to test our hypothesis that honeytokens could be used to mask the presence of devices that were transmitting a lot of data relative to other devices in the P2P network. First, we detail the physical setup of the simulation with several thousand shipping containers, each with an affixed P2P device. We then describe the placement and amount of deployed honeytokens. We perform several experimental simulations using various amounts of honeytokens. We conclude by describing the message set that was used to compare the systems.

A. System Setup

For our experiments, we simulated one ship with several thousand containers, each with a P2P sensor attached. As mentioned previously, the Emma Maersk is the largest container ship currently sailing [5]. It is capable of holding approximately 11,264 twenty foot equivalent containers. We located the dimensions of the ship and placed each of the 11,264 containers at a specific place on the ship. The ship is 397 meters in length, and 56 meters wide. Based upon the dimensions of the ship and other technical data that was obtained, the simulated containers were stacked 8 tall by 22 wide by 64 containers in length. We then simulated placing a sensor on the outside of each container, each having a specific location in three dimensional space.

B. Test Setup

Next, we simulated various amounts of honeytokens distributed throughout the ship. The honeytokens were placed randomly throughout the ship on the exterior of the containers and at the side of the ship's deck. We ran simulations with 50, 500, 1000, and 5000 honeytokens. We anticipated that if only a small amount of honeytokens were used then the produced noise would not be enough to effectively mask the actual communications. We compared these four scenarios against the system with no honeytokens.

The attacker only has a limited range in which they can receive signals from the sensors successfully. For our tests, we assumed that the attacker was at a point in the middle of the ship and that the attacker's receive antenna had a range of 50 meters [10]. In reality, an attacker could also be scanning from a port at the side of the boat, and we will take this into account in a future work. In order for the attacker to receive a transmission from a sensor or honeytoken, the receive antenna has to be less than 50 meters from the transmission source, or:

$$\sqrt{(S_x - I_x)^2 + (S_y - I_y)^2 + (S_z - I_z)^2} < 50 \text{ m}$$

where (S_x, S_y, S_z) represents the spatial coordinates of the attacker's antenna, and (I_x, I_y, I_z) represents the spatial coordinates of a particular sensor. The above equation gives the distance the object is from the antenna if a line is drawn from the sensor to the antenna in three dimensional space.

For all of the simulations, we ran the same series of simulated messages. For the simulations, the P2P sensors broadcast a system secure message every 60 minutes, unless the container had been breached, in which case a message is sent stating that fact. Also, the sensors wake every 10 minutes for 30 seconds in order to listen for breach signals or other transmissions from the system. We simulated the probability of each container needing to transmit a message at a given time slot as 5%, which is what we expected the probability to be in actual usage on the ship as the breach and other events are rare occurrences. We then simulated 2.5% of the sensors as "hot" sensors. These sensors communicated at a much higher rate, even though it was still a rare event for the system overall; for our experiments we used a Poisson distribution with a lambda value of one. Again, this relatively small amount of "hot" sensors is what was expected in actual usage. As the honeytokens only produced noise, we allowed them to all be considered "hot" sensors, with a Poisson distribution that was skewed towards a higher transmission probability. If we allowed the honeytokens to transmit all the time, this would easily be removed by the attacker as it would simply raise the noise floor threshold; instead, we allowed the honeytokens to act like the devices that were transmitting the most, which made the noise harder to remove as it looked like the real thing. We ran each simulation for a total of 3 hours, so as to collect a sufficient amount of data. At each ten minute interval when the sensors woke up, measurements were taken on the amount of simulated transmissions that the attacker would intercept given the constraints identified above. As we get more real world usage data, we will adjust these simulation properties to accurately reflect the environment.

V. RESULTS

We first ran the simulation with no honeytokens present to get a baseline comparison of how the system operates. As expected, there were certain peaks in the transmissions that were received which allowed the attacker to singulate out individuals or small groups of containers that were transmitting at a much higher relative rate compared to the rest of the containers. Figure 1 on the opposite column shows this weakness; in time frames four, twelve, and sixteen, the amount of communication is over twice what is normally transmitted. Time frames one, seven, thirteen, and nineteen are the highest communications where all sensors transmit a signal to indicate that they are still operational. As this is a system wide event, it does not reveal secure information.

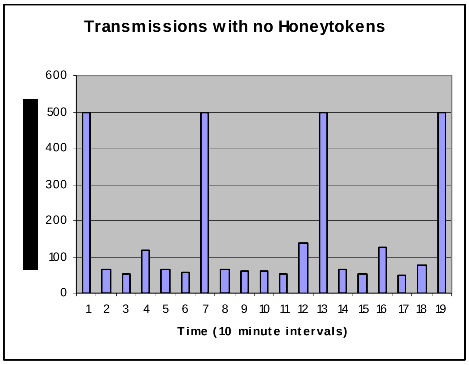


Figure 1: Intercepted transmission with no honeytokens.

As mentioned previously, our goal of using the honeytokens is to obfuscate the actual communications. Our first attempt at doing this used 50 honeytokens randomly distributed throughout the ship. While the honeytokens were still producing noise at the same rate as the most active sensors, they were unable to effectively mask the true signals. The results are presented in Figure 2 below, and it can be seen that the same peaks of communications that appeared in our baseline comparison in Figure 1 are still present.

We then proceeded to test the system with 500, 1000, and 5000 honeytokens. We expected to find that the intercepted communications appeared much more random as more honeytokens were added.

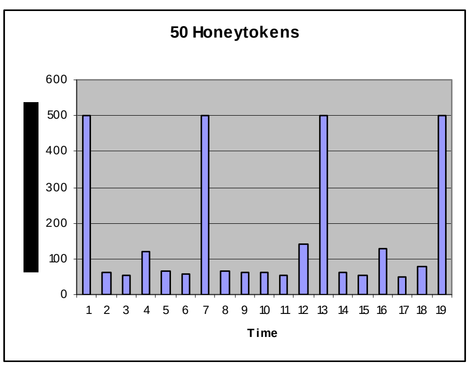


Figure 2: Intercepted signals with 50 honeytokens.

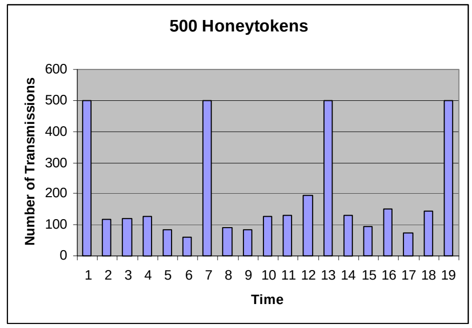


Figure 3: Simulation results with 500 honeytokens.

The results with the usage of 500 honeytokens are presented above in Figure 3. The communications are beginning to appear much more random; however the largest amounts of communications can still be seen at the fourth, twelfth, and sixteenth time frame. However, the other time frames have a much more similar distribution, and the peaks are beginning to even out.

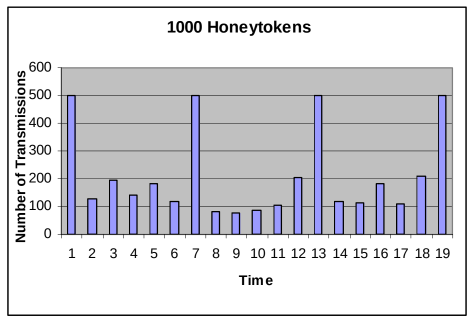


Figure 4: Results with 1000 honeytokens.

Figure 4 above shows the results when one thousand honeytokens were deployed randomly throughout the ship. At this point, the communications have become obfuscated. The honeytokens have masked the presence of the active devices by transmitting a sufficient amount of noise. While the clock signals that are transmitted once every hour are still present, this does not present much of a security risk as every device in the network transmits this.

Figure 5 shows the scenario when five thousand honeytokens were used; even the clock signals were masked by the honeytoken noise. The communications appear pseudo random, and an attacker can develop very little information about the system. In the case where five thousand honeytokens were used throughout the ship, the attacker is actually receiving significantly more noise than actual sensor communications, which makes actually deciphering any useful information difficult.

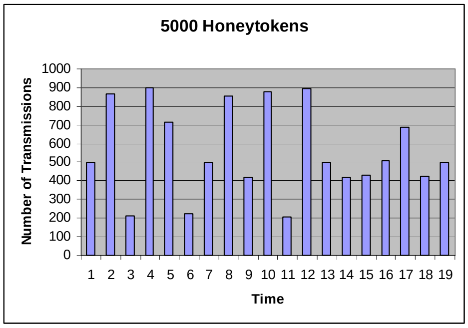


Figure 5: Results with 5000 honeytokens.

The greatest amount of obfuscation was observed when the largest amounts of honeytokens were employed, which is what was expected. However, it is likely not economically feasible to deploy such a great amount of honeytokens throughout the ship. When one thousand honeytokens were employed, sufficient noise was present to hide the actual communications from an attacker that might be listening in on and the one thousand honeytokens represent slightly less than ten percent of the number of sensors that are present on the ship. Also, in some circumstances, even less honeytokens might be appropriate.

VI. CONCLUSIONS AND FUTURE WORK

As pervasive computing become more common in our society, proper security measures must be in place that are aware of the physical limitations of the electronic devices [10]. In this work, we have focused on a maritime environment where wireless sensors are deployed onto shipping containers. In order to conserve battery, the sensors only transmit when required, saving battery. However, this opens up a risk in that an attacker can easily locate and disable a sensor that is transmitting a relatively large amount of data when compared to the other sensors in the collocated RF environment.

We have designed and tested a novel methodology that enables the mitigation of this risk. We have proposed deploying honeytokens throughout the ship. Honeytokens are devices that transmit realistic, but synthetic noise meant to distract the attacker and hide the actual valuable communications. The honeytokens transmit noise that appears realistic to an outsider; however it is easily separated and removed from the actual communications by trusted insiders. The honeytoken devices themselves are simpler than the sensors that are deployed on the shipping containers on the ship and they run on a separate power supply, conserving the battery life of the actual sensors.

In terms of future work, we would like to begin field trials of our work. We intend to measure more accurately the probabilities at which the sensors transmit data. We plan to test with physical devices and measure how far the range is of the devices.

We have simulated several scenarios with honeytokens deployed randomly throughout a ship. Our results showed that the communications were made the most random when the largest amount of honeytokens were used. In our tests, this occurred when 5000 honeytokens were deployed among the 11,264 containers. However, we also found that even a smaller amount of honeytokens are sufficient to remove the ability for an attacker to singulate out a sensor or group of sensors that are communicating consistently. When 1000 honeytokens were used, the communications appeared random to an attacker, effectively mitigating this risk. Our experiments showed how effective honeytokens are in this environment and our work was successful in identifying how and in what amount honeytokens should be used.

ACKNOWLEDGEMENT

This work is supported, in part, by research supported by the National Science Foundation (under Grant Nos. CNS-0619069, EPS-0701890, and OISE 0729792); NASA EPSCoR Arkansas Space Grant Consortium (#UALR 16804); and Acxiom Corporation (#281539).

REFERENCES

1. I. Mokube,, M. Adams, “Honeypots: concepts, approaches, challenges”, *Proceedings of the 45th annual southeast regional conference*, ACM Press, New York, NY, (2007), pp. 321–326.
2. C. McRae., R. Vaughn, “Phighting the Phisher: Using Web Bugs and Honeytokens to Investigate the Source of Phishing Attacks”, *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, (2007), pp. 270.
3. J. Yuill, M. Zappe, D. Denning, F. Feer, “Honeyfiles: Deceptive Files for Intrusion Detection”, *Proceedings of the 2004 IEEE Workshop on Information Assurance*, IEEE publishers, West Point, NY, (2004), pp. 116-122.
4. S. Weis, “Security parallels between people and pervasive devices”, *Proceeding of the 2nd IEEE International Workshop on Pervasive Computing and Communication Security - PerSec 2005,* IEEE publishers, West Point, NY, (2005) , pp. 105–109.
5. Maersk Line Facts and figures. Retrieved July 4, 2008 from the Maersk Line web site. http://www.maerskline.com/link/?page=brochure&path=/about_us/facts_and_figures.
6. Intermodal freight transport, Retrieved February 11, 2009, http://www.nationmaster.com/encyclopedia/Intermodalfreighttransport.
7. B. McCarty, "Botnets: Big and bigger," *IEEE Security & Privacy Magazine*, vol. 1, no. 4, July 2003.
8. R. Vogt, J. Aycock, and M. Jacobson, "Army of botnets," in *Proceedings of 14th Annual Network and Distributed System Security Symposium (NDSS)*, February, 2007.
9. E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim, "A survey and comparison of peer-to-peer overlay network schemes," *IEEE Communications Surveys and Tutorials*, vol. 7, no. 2, 2005.
10. S. Weis, “Security parallels between people and pervasive devices”, *Proceeding of the 2nd IEEE International Workshop on Pervasive Computing and Communication Security - PerSec 2005,* IEEE publishers, West Point, NY, (2005) , pp. 105–109.
11. J. White, “Radio Frequency Identification (RFID) Honeytokens as a Mitigation Tool against Illicit Inventorying”, in *Proc. International Conference on Information Security and Privacy (ISP-2009)*, Orlando, Florida, July 13 – 16, 2009, pp. 67 – 73
12. S. Geoghegan, G. McCorkle, C. Robinson, G. Fundyler, S. Ramaswamy, M. Tudoreanu, R. Seker, J. Brown, and M. Itmi, “A Multi-Agent System Architecture for Cooperative Maritime Networks”, in *Proceedings of ISC 2009*, Toronto, Canada, 2009.

Integrating Information Security into Quality Management Systems

Dr. PhD Margareth Stoll Margareth.stoll@eurac.edu
EURAC Research, Drususallee, 1
39100 Bozen, South Tyrol, Italy

***Abstract*-Due to globalization, stronger competition, increased complexity, information explosion, interconnection and extensive use of technology, information management is a main performance driver and key differentiator for sustainable organization success. Data, information, knowledge and asset are exposed to most different threats. Therefore more than 5800 organizations worldwide are implementing an information security management system in accordance with ISO/IEC 27001. Many organizations of different sizes are implementing also standard based management systems, such as ISO9001 or others. On the basis of several practical applications in most different organizations are described applied methods and experiences by implementing suitably holistic management systems for risk prevention and quality improvement. This promotes effectiveness, efficiency, legal conformity, information security awareness and organization development including improved information security for sustainable organization success.**

Keywords: information security, quality management, holistic enterprise management, data protection, risk management

I. Introduction

A Starting Situation

Due to globalization, ever stronger competition, information explosion, increasing interconnection, extensive use of IT-systems and the growing complexity data and information management for knowledge generation and continual improvement are main performance driver and key differentiators for competitive advantages and sustainable organization success. Thus data, information, knowledge, IT and production systems become a completely new role. But data, information and knowledge are exposed to most different threats, such as physical and environmental threats, technical threats, organizational and human related threats. The Computer Security Institute (CSI) and the FBI describe that virus attacks are the source of the greatest financial losses, followed by unauthorized access to networks, lost or stolen laptops or mobile hardware, and theft of proprietary information are the next and all together was reported a lost account for more than $39,370,717 [1]. The kes Microsoft study and the report about the IT security situation in Germany explain that collaborators error and negligence are the main cause for damage followed by software and hardware problems [2], [3]. In the last years the data protection law requirements are sharpened, the IT governance requirements and thereby data integrity requirements are increased and based on the main role of data, information and knowledge the requirements for confidentiality and availability are increased.

In this respect information and system availability, confidentiality and data integrity, as well as data protection and fulfillment of legal and regulatory requirements is central for organizations' success.

One million organizations of different sizes and scopes are implementing, already since several years, quality management systems, such as ISO9001, or other management systems based on international standards (e.g. environment ISO14001, IT service management ISO22000). All these management systems require common principles: organization objectives and strategies, business processes, resource management and continuously optimization of the organization [4]. The established management system must be documented, communicated, implemented and continuously improved. These systems are implemented more frequently holistic, whereby are integrated according with the organizational purpose and objectives different aspects, like quality, environment, hygiene, occupational health and safety, as well as personnel development, resource management, IT - management, communication management, controlling and also knowledge management.

B Purpose of the article

The increased requirements for information security including availability, confidentiality and integrity, international standards such as IEC/ISO 27001 [5], IEC/ISO 27002 [6], ISO/IEC 20000-1, Sarbanes-Oxley or other legal or regulatory requirements, the improved risk awareness promote a holistic information security approach. There are just now more than 5800 IEC/ISO 27001 information security systems implemented and certificated [7]. The information security management systems are often implemented parallel to other standard based management systems. The trend to holistic management systems and the need of developing an integrated solution to support optimally all requirements of an organization and to promote a consistent and integrated implementation and operation influenced us to integrate an information security management system with an ISO 9001 quality management system. This will enrich management systems by information security aspects, use common synergies and resources, and it will contribute efficiently, re-

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_79,

sources carefully, fast and successfully to support information security for sustainable organizational development and success.

C Research Approach

We developed a holistic information security management model by integrating the ISO 9001 standard for quality management systems or other international standards for management systems and the information security management standard in accordance to IEC/ISO 27001 and IEC/ISO 27002. This model was implemented after in different small and medium sized organizations. The case study results were collected by established process and system measurement methods and interviewing the managers and collaborators.

D Structure of the Article

Firstly we describe based on the starting situation the project objectives [II]. After we explain the requirements for our approach based on the main requirements of quality management systems (ISO 9001) and other international standards for management systems [III A]. Also the main requirements for information security management in accordance to IEC/ISO 27001 and IEC/ISO 27002 [III B] are illustrated. After we present our holistic information security management model [IV] and the implementation of that model [V] with the development of policy and objectives [V A], the risk assessment, risk treatment and business continuity plan [V B], the process analysis and process improvement [V C], the resource management [V D], the continually improvement of the whole system and the system documentation [E]. Finally we report about our project experiences and results of the implementation in different small and medium sized organizations [VI] with the achievement of the project principles [VI A] and we reflect about success factors [VI B]. At the end we give an outlook [VII] and our conclusions [VIII].

II. PROJECT OBJECTIVES

How can be integrated the requirements of information security management in accordance to IEC/ISO 27001 and IEC/ISO 27002 and the requirements ISO 9001 quality management systems or other standards based management systems in order to contribute to organizations' success and to be considered, used and improved as base for organization development?
By integrating information security management in accordance to IEC/ISO 27001 and IEC/ISO 27002 into a, the organization best adapted, holistic information security management model we expect to foster:

- effectiveness,
- efficiency and cost reduction,
- shorter initial training periods for new collaborators, and
- improved information security.

Thus information security and thereby availability of data, documents, asset and infrastructures, confidentiality of information, documents and data, as well as data integrity should be optimized and the organization will be promoted in achieving its objectives including information security objectives and reducing risks.

III. MAIN STANDARD REQUIREMENTS

A Main Requirements of Standard Based Management Systems

The ISO 9001 quality management standard and other standards for management systems require common principles [4]:

- Organizations' objectives and strategies must be established regarding stakeholder requirements [see top of Fig.1].
- All business processes including the management process, support processes, resource processes and optimization processes must be defined and promote the optimized fulfilment of the organizations' objectives under the focus of the respective standard [see the graphic under the top of Fig. 1].
- Process oriented resource management must be promoted including human resource development, IT – management and other infrastructures, tools and instruments [see bottom of Fig. 1].
- The organization, their objectives and strategies, services/ products and processes must be continually optimized according to established processes in sense of a PDCA cycle (plan, do, check, act) [see the circle around in Fig.1].

The established management system must be structured and systematically documented and communicated within the organization and the collaborators must be continually motivated for implementing the system and for recommending improvements.

These standard based management systems are implemented more frequently in a holistic way. In accordance with the organizational purpose and objectives are integrated into a management system eventually environmental, hygienic, occupational health and safety aspects, as well as human resource development, resource management, knowledge management, IT - management or controlling.

Fig. 1. Main requirements of standard based management systems

B Main Requirements of Information Security Management Systems

We present the requirements of ISO/IEC 27001 [5] and ISO/IEC 27002 [6]:

- An information security policy must be defined and approved by the management.
- A risk assessment must be conducted to establish a risk treatment plan to reduce risks to acceptable levels of risk. For the identified remaining risks a business continuity plan must be developed, implemented, maintained, tested and updated regularly.
- The needed resources must be determined and provided. All collaborators must be competent to perform their tasks and be aware of the relevance and importance of their information security activities and how they can contribute to the achievement of the information security management system objectives.
- The effectiveness and adequacy of the information security management system must be continually improved using controls, objectives, audit results, by monitoring, applying corrective and preventive actions and management review in sense of a PDCA cycle (plan, do, check / study, act).

The established information security management system must be structured and systematically documented, communicated, implemented and continual improved.

IV. HOLISTIC INFORMATION SECURITY MODEL

Due to the great diffusion of ISO 9001 quality management systems or other standard based process oriented management systems, the role of knowledge and information as main performance driver and key differentiator and the various most different threats for information, data and asset, as well as the alignment of different standard based management systems we integrate an information security management system and standard based management systems to a holistic information security management model.

Based on the main requirements of standard based management systems [III A] and the information security management system requirements [III B] we developed a holistic information security management model:

- The corporate vision, policy, objectives and strategies are extended by information security aspects [see top of Fig. 2].
- A risk assessment must be conducted to establish a risk treatment plan to reduce risks to an acceptable level of risk. For the identified remaining risks a business continuity plan must be developed, implemented, maintained, tested and updated regularly [see graphic under the top of Fig. 2].
- All business processes are analyzed regarding also information security aspects and all in the risk assessment identified risk treatments and the business continuity plan are

Fig. 2. The Holistic Information Security Management Model

integrated into the operational processes and thereby implemented, maintained, tested and updated regularly to ensure there effectiveness. Afterwards the business processes are improved according to the established objectives. Thereby all stakeholder requirements inclusive the information security aspects are improved. The process description establishes for all process steps the associated responsible and accountable function and the relevant information security requirements (corporate, regulatory or legal requirements) [see middle of Fig. 2: the main business processes start from the first contact with the customer to the delivery of the product/service and the smile of the customer].

- The resource management specifies competence objectives aligned to the corporate objectives for all functions. Appropriate trainings or other actions must be taken to achieve and maintain these objectives according to defined processes. Their effectiveness must be evaluated. Thereby also all required information security competences and the collaborators' awareness is constantly identified and improved. The organization must also define, plan and provide the necessary resources, tools and instruments to obtain and continually improve its objectives including information security objectives [see bottom of Fig. 2].
- The effectiveness and adequacy of the established holistic management system and objectives must be evaluated periodically by suitable methods for monitoring and, where applicable, measurement. It must be continually improved according to established processes in sense of a PDCA cycle (plan, do, check / study, act) [see the circle in Fig.2].

V. IMPLEMENTATION

A Policy and Objectives

Starting from the requirements of interested parties, legal or regulatory requirements, contractual security obligations, the characteristics of the business, the organization, its location, environment, assets and technology we elaborate using Quality Function Deployment [8] the corporate policy including information security elements and integrate it to a holistic corporate policy. Thereby we define the information security management system scope and boundaries. We deduce after from the policy the objectives and strategies inclusive information security objectives, such as availability, confidentiality and integrity. Also all legal, regulatory and corporate information security requirements are analyzed and adequate information security strategies deduced. Thus the entire organization is focused on the fulfillment of stakeholder requirements including information security objectives and requirements.

B Risk Assessment, Risk Treatment and Business Continuity Plan

A risk assessment is conducted to establish a risk treatment plan to reduce risks to an acceptable level of risk. Therefore we identify the assets, define the required levels of confidentiality, integrity and availability for the assets, identify the potential threats to those assets, define the realistic likelihood of security failures, the impacts that losses of confidentiality, integrity and availability may have on the assets, the controls currently implemented, estimate the levels of risks regarding the implemented controls and elaborate for the higher risks adequate risk treatment plans. The management must approve whether the risks are acceptable or further risk treatments must be implemented. For the risk treatments appropriate risk control objectives and controls are selected and approved. The risk treatment plan is integrated into the operational processes. Thereby the risk treatment and the existing standard based management system are integrated. For the identified remaining risks a business continuity plan is developed, integrated into the operational processes. Thus it is implemented, maintained, tested and updated regularly to ensure that it is effective to maintain or restore operations and ensure availability at the required level and in the required time scale following interruption to, or failure of, critical business processes.

C Process Analysis and Process Improvement

The management process, all business processes including supporting processes, resources processes and optimization processes are analyzed bottom up by interviewing the concerned collaborators and defining necessary information, data, knowledge and documents [9]. Thereby also all relevant standards, guidelines, legal and regulatory requirements are defined and integrated. The processes are optimized regarding the established organizational objectives, as well efficiency, effectiveness, information security requirements, data protection and other corporate, legal and regulatory requirements. Thereby procedures, regulations, legal and regulatory interpretations, checklist, forms and IT systems are harmonized and also knowledge and information management improved [9], [10].

By process modeling for all process steps the responsible and accountable functions, as well as necessary documents, information, tools, IT - applications including the observing information security laws and regulations are defined. For all documents we deduce the responsible function for establishing, generating, changing, handling and the accountable for it, as well as the functions, which will collaborate and/or be informed. Also external information sources or receivers (document logistic), the data protection class in accordance to the data protection law and further the required archiving methods (for current, intermediate and historical archives) are identified. Thereby we establish necessary and licit access rights, archiving procedures, information security requirements regarding confidentiality, availability and integrity, necessary data encryption with encryption procedures, as well as signature rights and signature procedures as basis for the implementation of adequate signature regulations. All treated data are examined for there necessity and lawfulness. Thus all information flows, documents, data and document handling and all processes are analyzed regarding information security and optimized.

Function profiles and required competences are deduced from the role of the different functions at the single process steps.

D Resource Management

The organization must determine and provide due to standard requirements necessary resources, tools and instruments to obtain the established objectives and to continually improve the organization. Thereby an optimal infrastructure for information security is promoted. Training and competence objectives are planed and implemented according to defined processes and their effectiveness are evaluated. Thus also the competences and awareness of all collaborators and persons to whom are assigned information security responsibilities are promoted systematically and structured, their effectiveness evaluated and possibly necessary corrective actions are taken.

E Continually Improvement

Due to standard requirements we deduce from the established corporate objectives including information security objectives the business process objectives with corresponding measurement methods to demonstrate the process ability to achieve planned results. When planned results are not achieved, corrective and eventually preventive actions must be taken. All corrective actions, improvements or changes must be discussed, approved, documented, communicated, implemented and their effectiveness evaluated. Corrective actions, improvements or preventive actions are introduced also by means of collaborator ideas, the results of periodically internal and external audits, feedbacks from stakeholders or external sources (e.g. supplier, technical journals). Thereby also information security is constantly evaluated and if necessary changed and optimized. Organizational knowledge is changed and therefore new individual learning becomes possible. Thus the knowledge and learning spiral is constantly pushed again and the organizational knowledge base extended [12].

F System Documentation

The established management system with policy, objectives, strategies, processes, risk assessment and risk treatment plan, business continuity plan, function profiles, templates, checklists and others must be documented. Every collaborator must be trained on it. All changes of the documentation must be distributed traceably and trained. The collaborators and particularly the managers must implement constantly the established processes and procedures.

To promote need-oriented, workplace integrated access to system documentation we prepare the documentation regarding didactical principles [12], [13] and distribute it electronically through web-based intranets, pdf or organizational learning and knowledge systems based on constructivist theory [14].

VI. Project Experiences and Results

This holistic information security model was implemented with distinct process oriented standard based management systems in different small and medium sized organizations. Thereby implementing process thinking, harmonizing and simplifying processes, process controlling and overall maintain sustainable security awareness were great challenges.

A Achieving the Project Objectives

The following case study results were collected by established process and system measurement methods and interviewing managers and collaborators:

- Effectiveness: The fulfilment of the established objectives is periodically controlled and studied by defined measurement methods and if necessary there are due to standard requirements elaborated and implemented appropriate corrective or prevention actions. Their effectiveness is controlled. Thereby the enterprises achieve their planned objectives in average more than 92%. In a service organization e.g., where confidentiality was a main factor for customer satisfaction it was improved constantly from 85% at the end of 2002 to 92% at the end of 2008.
- Efficiency and cost reduction: Integrating information security into an existing standard based management system reduces in our case studies the effort for the development about 10-20% and thereby also the costs. The advantages are still higher during the implementation, where only the audit is extended and the risk assessment must be maintained.
- Shorter initial training periods for new collaborators: New collaborators are quickly introduced into the management system at their first working day. Thereby they focus on the principle information. Afterwards they access the documentation work-integrated and need-oriented, when they have questions for fulfilling their job. The lead time could be abbreviated around a half. New collaborators execute their tasks faster well regarding information security, whereby the productivity increase and possible errors are avoided.
- Information security improvement by employee involvement: Due to the standard all ideas, optimizations and suggestions of the collaborators must be evaluated, eventually implemented systematically and structured and the effectiveness of the implementation must be controlled. Thus information security is promoted. Thereby we received averaged monthly 0.1 until 0.6 ideas from each collaborator. This continuously organizational improvement by ideas and suggestions of the collaborators and the systematic und structured process for determining, implementing and evaluating the effectiveness improve information security and the organization itself.

Data protection and information security legal and regulatory requirements are analyzed, implemented and maintained by defined improvement processes. Thereby the legal and regulatory conformity is sustainable implemented. This is specially appreciated by managers, controllers, risk managers and/or responsible persons for data protection and/or information security.

All the discussions about information security during the system introduction and the trainings increased severely the information security awareness of the collaborators.

By integrating information security completely into a standard based management system the information security management system uses synergies, such as the existing process and system thinking, the systematic process oriented approach, the existing policy, strategy, objectives, the training and improvement processes, and the management responsibility. In many technical organizations the collaborators feel the information security management system more helpfully than the quality management system or others and thus also the existing standard based management system benefits from the integration. The collaborators feel the integration more than an extension and enriching.

B Success Factors

Corporate culture processes and technology must be integrated optimally according to the organizations' objectives and to collaborators needs and requirements, as well as considering didactical principles. Standards, methods and technology are thereby only tools, which support organizational development and information security so far as this is admitted by the culture. Therefore we need an open, confident based, participative corporate learning culture with criticism and change readiness [11].

The collaborators should be interested in new knowledge, able for self-driven learning, have personal employment, team ability and change willingness [11]. All managers must promote constantly information security and motivate their collaborators in following these principles.

Introducing information security into holistic standard based management systems requires a, the organization best adapted, system. This certainly can not be achieved by using a purchased general manual, which does not correspond with the lived processes and corporate objectives. Further the management system must be constantly and effectively implemented.

This holistic information security model requires from the system or information security manager, a part from standard management system skills, also information security skills, the acquisition of necessary change management, organizational learning, and controlling knowledge, the knowledge about all relevant information security and data protection laws, standards and regulations, as well as experience in information technology, knowledge management and information management.

Sufficient IT-infrastructure and IT-support are also very important for the project success. Only an appropriate IT system, technical equipment and infrastructure in accordance with organizational objectives promote continuously optimization of information security and sustainable organizations' success.

VII. Outlook

Due to these excellent project experiences in several organizations with different management systems there should be introduced enhanced information security management into standard based management systems regarding all success factors [VI B].

Standard management system trainings should inform about the principles of information security management systems and vice versa.

The management systems standards should emphasize the importance of information security and require the integration of information security into management systems.

VIII. Conclusion

Integrating information security management into a process oriented standard based management system enriches it. Thereby information security can be introduced efficiently, resources carefully, fast and successfully using many common synergies.

Especially in organizations, where data, information and knowledge or availability of services, asset and/or infrastructure or confidentiality or integrity has a high importance for organizations' success, information security management should be enhanced by integrating it into all existing management systems.

Implementing the holistic information security management model promotes information security and thereby in accordance to organizations' objectives availability, confidentiality and integrity, effectiveness, efficiency, legal and regulatory conformity, information security awareness and organizational development including information security improvement to guarantee stakeholder orientation and sustainable organizations' success.

References

[1] Gordon, M. Loeb, W. Lucyshyn: *CSI/FBI Computer Crime and Security Survey*, i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2006, 2006

[2] <kes>-Microsoft® Sicherheitsstudie 2006, kes, vol. 4-6, Sonderdruck, 2006

[3] Federal Office for Information Security: *Report about the IT security situation in Germany*, bsi.de/english/publications/securitysituation/-Lagebericht_2007_englisch.pdf

[4] ISO, ISO 9001:2008, Quality Managemnt Systems – requirements, ISO, 1.12.2008.

[5] ISO, *ISO/IEC 27001:2005*, Information technology – Security techniques – Information security management systems – Requirements, 5.10.2005.

[6] ISO, *ISO/IEC 27002:2005*, Information technology – Security techniques – Code of practice for information security management, 15.6.2005.

[7] http://www.iso27001certificates.com/

[8] Y. Akao, *Quality Function Deployment*, integrating customer requirements into product design, Productivity Press, Portland, 1990.

[9] M. Stoll *Managementsysteme und Prozessorientiertes Wissensmanagement* in Proc. 4th Conference on Professional Knowledge Management – Experiences and Visions, Gito Verlag, Berlin (2007), volume 1, pp. 433-434.

[10] T. Davenport: Thinking for a living, how to get better performance and results from knowledge workers, Harvard Business School Press, Boston, 2005.

[11] P. Senge: Taking personal change seriously, The impact of Organizational Learning on management practice, Academy of Management Executive, vol. 17, p 47-50, May2003.

[12] M. Rosenberg: E-Learning, Strategies for Delivering Knowledge in the Digital Age, McGraw-Hill, New York, 2001

[13] G. Reinmann-Rothmeier Mediendidaktik und Wissensmanagement. in MedienPädagogik 10, www.medienpaed.com/02-2/reinmann1.pdf, 2002.

[14] M. Stoll, *Workplace Process Integrated Learning and Knowledge Organization*, in H. Maurer and K. Tochtermann Eds. Proc. I-Know 07, 7th International Conference on Knowledge Management, J.UCS Journal of Universal Computer Science, Graz, 2007.

Evaluation of Asynchronous Event Mechanisms for Browser-based Real-time Communication Integration

Keith Griffin
Cisco
Galway, Ireland
kegriffi@cisco.com

Colin Flanagan
University of Limerick
Limerick, Ireland
colin.flanagan@ul.ie

Abstract — **Desktop based real-time communication applications are commonly used for presence based instant messaging and telephony applications. Such applications use installed desktop components to handle real-time asynchronous events on the client originating from the communication system. When the client is based in a web browser these installed components are not available however browser-based communication applications are required to handle the same type of asynchronous events. Moreover, browser-based clients which typically run over HTTP are challenged by HTTP itself which is designed to be a synchronous request-response protocol. We contend that a suitable mechanism can be found to deliver asynchronous real-time events to browser-based applications**

Keywords: Browser, HTTP, Communication Integration, Telephony, Presence, Web.

I. INTRODUCTION

In real-time communication integration systems such as those providing telephony, instant messaging and presence integration asynchronous events are a core and mandatory feature. Basic events like a presence state change from busy to available or lifting a telephone handset resulting in an offhook event being delivered to an application are fundamental to traditional desktop based communication integration applications. An example of such an application displaying presence state is shown in Fig. 1 [1]. Presence is defined as a person's availability and propensity to communicate [2]. A person's availability can be reflected by a communication state. This state is often presented as a colour, icon or colour coded icon. A red phone icon for a contact may represent that the contact is busy and unavailable for voice communication, with amber representing that they have been idle for some time and green representing that they are available for voice communication. In this case the red, amber and green scheme which is commonly used in road traffic signals is being use to represent availability and a phone icon is being used to represent telephony or voice communication. In Fig. 1 the scheme that is used is to overlay a graphical indication along with a colour scheme. A positive tick mark on a green icon is used to display availability while an icon representing the hands of a clock is overlaid on a blue icon to represent an inactive state.

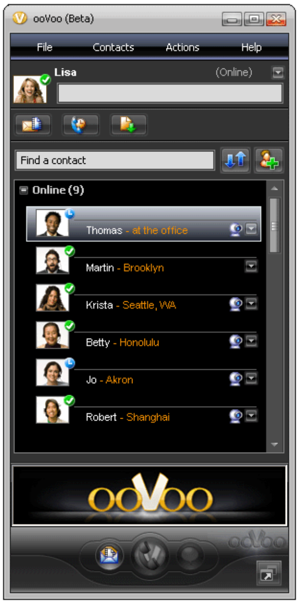


Fig. 1 Presence enabled buddy list in a desktop communication client

A user's presence state can be affected by multiple sources as described in Fig. 2. These sources result in a mix of user and system driven presence state changes.

- User driven presence: A user manually sets their presence or availability state using a user interface such as a drop down menu. Presence states are often restricted to include "available", "unavailable" or "busy" and "idle".
- System driven presence: A user's state is automatically set by a system event e.g. logging off from an instant messaging client would automatically set the user to "unavailable".

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_80, © Springer Science+Business Media B.V. 2010

Presence Source	Sample Presence States
User	Busy, Available, Idle, In a meeting
System	Logged in, Logged out, Idle
Calendar	Available, Busy
Online Meeting System / Web Share	In a meeting, Meeting state
Telephony	One the phone, call state
Instant Messaging	Message State

Fig. 2 Sample input sources for presence

Regardless of how a presence state change occurs the change must be reflected in applications subscribing to that user's presence in real-time. Some applications can integrate directly with a protocol while others use middleware based integration [3]. While it is clear that presence functionality is available to middleware and installed desktop applications, it is proposed that suitable mechanisms can be found to offer presence functionality to browser-based applications. We will use presence as a sample real-time communication integration service to discuss various mechanisms to extend real-time asynchronous presence events to a browser-based application.

II. Asynchronous Events for Browser-based Applications Overview

Browser-based communication applications typically rely on HTTP as the protocol used to communicate with a web server. HTTP 1.0 [4] improved on previously defined versions of the protocol which provided for raw data transfer over the internet. Improvements included allowing messages to be of MIME like format containing meta information about the data being transferred. Amongst other things HTTP 1.0 did not sufficiently take into consideration persistent connections. However persistent connections are considered in HTTP 1.1 [5] in order to reduce the number of TCP connections and reduce congestion. However this does not address the ability for a client to receive real-time asynchronous events from a server to a browser-based application. An asynchronous event effectively amounts to a response without an explicit request. The browser-based application will have implicitly solicited an asynchronous event as part of an earlier request but the timing for an availability of that event will be unknown and driven by other events in the communication system. A mechanism is required which will allow the delivery of asynchronous events which occur in the communication integration system to browser-based clients. However any such mechanism must respect the behavior of web-based systems and not allow unsolicited events from the server to the client which could create a security vulnerability.

One way to handle asynchronous events would be to use a continuous polling mechanism but polling would impact both the performance and scalability of the system which are among the benefits that HTTP presents in the first place. It has been found that several alternatives exist to continuous polling which will be explored in the following sections.

III. Overview of Alternative Approaches

Finding a solution for asynchronous events to browser-based applications is not a trivial task. Browser-based applications typically run over HTTP which despite having persistent connections is designed as a synchronous, stateless and connectionless protocol. Furthermore communication related services such as presence generate asynchronous events which require real-time handling. Take the example of a user who has a presence state of "BUSY - ON THE PHONE" as a result of an integration between a telephony and presence service. This busy state may be displayed on a number of other clients who are subscribers or followers of this user's presence state. The number of clients depends on the number of others who subscribe to the user's presence and whether those users have a running client capable of displaying the presence state. When the user finishes the phone call the telephony system will update the presence system to reflect a state of "AVAILABLE". Once this state change happens, all other clients subscribed to that user's presence must be updated in real-time. Without this real-time update the value of the system would be significantly diminished. In the case of an installed desktop client which typically includes a local protocol handler, receiving an asynchronous update could be described as straightforward although not trivial as it is still a complex task. It is straightforward as the locally installed desktop component can include a protocol handler acting as a listener and ready to receive and handle incoming asynchronous events over a persistent connection. In the case of a browser-based client it is not so straightforward. The application running in the browser can make a request for a particular user's presence from a presence web service and it will get a response with the presence state. If that user's presence state changes at an unknown time in the future how will the browser-based application become aware of this change? In the following sections we will explore mechanisms to extend asynchronous real-time events to browser-based presence applications.

A. *Continuous Polling and Streaming*

One straightforward way to keep the client aware of presence updates on the server is to not use an event mechanism at all but instead to continuously ask what the state is. Using this approach, rather than waiting for the server to push events to the browser-based application the application implements a continuous cycle of HTTP requests for a given user or list of user's presence. For example request a presence state every second, receive and immediate response, close the connection and repeat. Various approaches to this technique have been suggested and implemented from the brute force technique suggested above to more elegant standards submissions such as WS-Polling [6]. An obvious disadvantage of continuous polling is that it results in the server handling all of the polling requests and as a result can be very inefficient on the server and also on the client which has to handle continuous page refreshes. An improvement to continuous polling would be to use a technique commonly referred to as streaming [7]. This involves making a request from the browser-based application to the presence web service and then not closing the connection at the server after the response has been sent. Subsequent asynchronous presence events for the initially requested user would then be streamed over this connection. Unlike continuous polling the server is not required to handle continuous requests however it is required to keep a connection continuously open which means that this connection cannot be used for other purposes. This approach takes the connectionless, synchronous request-response nature of HTTP and turns it into a connection oriented approach like an installed desktop client might use with a persistent TCP socket connection. For a web based application providing real-time asynchronous events this is not thought to be a scalable solution, as it locks up too many connections. In particular, browsers often place a limit on the number of simultaneous connections they will hold open to a server.

A technique that can be combined with polling for a more efficient approach is Asynchronous Javascript And Xml (AJAX). Ajax consists of a number of technologies that combine as a mechanism which can be used to poll for a specific update without updating the entire content of the page. Two key technologies used by Ajax are JavaScript and XMLHttpRequest (XHR) [8]. XHR can be used to request or poll a presence web service for updates. Following such an update the XHR object can use a call back function to update the browser-based application using JavaScript. An advantage of Ajax is that it can be used to update a specific data set such a user's presence state without requiring an entire page refresh. This offers a more efficient approach however it still relies on a brute force polling technique to realize presence state changes on the server, as described in this example.

B. Long Polling

Long polling is a technique whereby a browser-based client sends a request to a server where the server holds the request open until a response is available. This is similar to the streaming technique in that the request is held open but different in that the connection can be closed either due to a defined timeout or when a response is received. Once a connection is closed a new connection is initiated. As a result a given client typically has a request open when an asynchronous event such as a presence event becomes available.

The Bayeux Protocol and Comet

Bayeux is a protocol used to transport messages between a web server and web browser using named channels [9]. Messages can be delivered as follows:

- Client to server
- Server to client
- Client to client (via server)

Bayeux suggests that all requests originate from the client i.e. a server may not initiate a connection to a client unless a client has previously sent a request to a server which supports the browser security model. However asynchronous events intended for a given client are supported once a client has made a request to the server.

The transport for server to client event delivery can terminate the HTTP response after all messages are sent as in the case of a polling transport or use a streaming technique that allows multiple event messages to be sent in the same HTTP response. Two transport types are supported by Bayeux: Long-Polling and Callback-Polling. Callback-polling is a recommended but not mandatory transport type as defined by Bayeux. Callback polling uses HTTP GET to pass the name of the callback to be triggered to the server whereas long polling uses a POST request and response.

Connections between the client and server are negotiated with handshake messages that allow the content type, authentication, protocol version and other parameters to be agreed. Comet is a term often used to describe the behavior of the Bayeux protocol. In fact the Bayeux protocol definition was led by the same person that coined the term "Comet". Cometd is an implementation of the Bayeux protocol by the Dojo Foundation [10].

Bi-directional Asynchronous Events with LongPolling

In order to support bi-directional communications as required for asynchronous presence updates, Bayeux clients use two connections to a server to allow client to server and server to client messaging occur simultaneously as shown in Fig. 3. Bayeux offers a transport option called *long-polling* which attempts to hold open a request until there are related events to deliver to the client. The intent is to always have a pending request to deliver events as they occur. The reconnect and interval advice fields are used to maintain requests over time and address resource starvation and increased server load. In the worst case this behavior can degenerate to that of traditional polling but used efficiently it can minimize both latency in server to client message delivery and the network resources required for the connection.

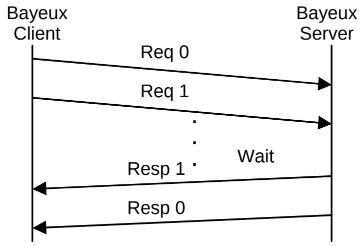


Fig. 3 Bi-directional Asynchronous Operation

Bidirectional Streams Over Synchronous HTTP

The XMPP Standards Foundation describes Bidirectional streams Over Synchronous HTTP (BOSH) [11] as *"a transport protocol that emulates the semantics of a long-lived, bidirectional TCP connection between two* *entities (such as a client and a server) by efficiently using multiple synchronous HTTP request/response pairs without requiring the use of frequent polling or chunked responses".* As can be seen from this definition, BOSH is a long polling technique and claims to be more efficient in terms of bandwidth and responsiveness than Ajax. The BOSH standard calls for the use of a HTTP Connection Manager which is effectively a HTTP web server capable of implementing the long polling technique as described in the standard. BOSH works by having the HTTP connection manager respond to a browser side request only when a server side event is available. Once the client has received this event it immediately makes another request to provide a vehicle for the next asynchronous event. The differences between the BOSH and Bayeux/Comet techniques appear to be minimal and are related to implementation differences when run over HTTP 1.0 and 1.1.

Interestingly, a standard has been defined for the use of the commonly used Extensible Messaging and Presence Protocol (XMPP) [12] standard with BOSH. The XMPP over BOSH standard [13] defines how XMPP stanzas can be transported over HTTP. As defined in XMPP an XML stanza is a discrete unit of structured information which can be used to represent a user's presence state. Clearly this proven technique provides a mechanism for real-time asynchronous presence events to be delivered to browser-based applications for XMPP based presence and instant messaging systems.

C. *Web Sockets*

Two key components of the HTML5 [14] specification, as it relates to real-time asynchronous events for browser-based applications, are the web sockets API [15] and *server sent events* [16]. Web sockets attempt *"to enable Web applications to maintain bidirectional communications with server-side processes"*. The web socket protocol [17] claims to provide an alternative to HTTP polling for bi-directional communications between a web page and server by using a single TCP connection to manage traffic in both directions. It uses HTTP to achieve a handshake between client and server by requesting a HTTP "upgrade: websocket". Following this the underlying TCP connection is used for bi-directional client-server communications. If the protocol delivers on this claim it will offer a single socket connection capable of bi-directional messaging between client and server. This sounds like the definition that one might expect for an installed desktop client running natively over TCP.

Once a websocket is established server sent events can be used to push real-time asynchronous events to a browser-based application. This is achieved by using the event source interface which allows for the creation of an event source object in the client and registering an event listener.

IV. ASYNCHRONOUS EVENTS FOR BROWSER-BASED COMMUNICATION INTEGRATION

Two of the previously outlined mechanisms were used to generate results for presence updates in a browser-based application, namely continuous polling and BOSH. Ajax was used as a common technique to update only the presence data-set in the browser without requiring a full page refresh.

A. *Results for Continuous Polling*

Using Cisco's Presence Web Service [18] a browser-based client was used to get and set presence states. In this case presence updates were polled using Ajax at the client. The resulting presence state change was observed in the browser client as shown in Fig. 4. The web page shows a client for user Colin Flanagan who is subscribed to presence state changes for user Keith Griffin. The presence state for Keith Griffin has just changed to Do Not Disturb (DND) at this point of the test.

Fig. 4 Presence enabled client with continuous polling

While performance related timing information was available in the traces the goal of the test was not to measure performance but to determine if asynchronous presence state changes could be successfully delivered to the browser. Trace output in Fig. 5 shows a basic presence update delivered to the user Colin Flanagan as a result of user Keith Griffin changing presence to a "Do Not Disturb" (DND) state.

```
"<presenceList type="basic">[\n]"
httpclient.wire.content - <<
"<contact contactURI="griffin@cisco.com"
presenceStatus="DND"/>[\n]"
"</presenceList>[\n]"
```

Fig. 5 Trace displaying XML Presence Representation for DND

B. *Results for Long Polling using BOSH*

Using the sample JavaScript libraries with the Jabber Extensible Communications Platform (XCP) XMPP based implementation [19], BOSH long polling was observed in the client shown in Fig. 1. A long-poll interval set to 30 seconds was used. When an event was received asynchronously on the server it was sent to the browser as a response to an open HTTP POST request. Once the event was received in the browser a new HTTP POST was initiated according to the BOSH long polling technique.

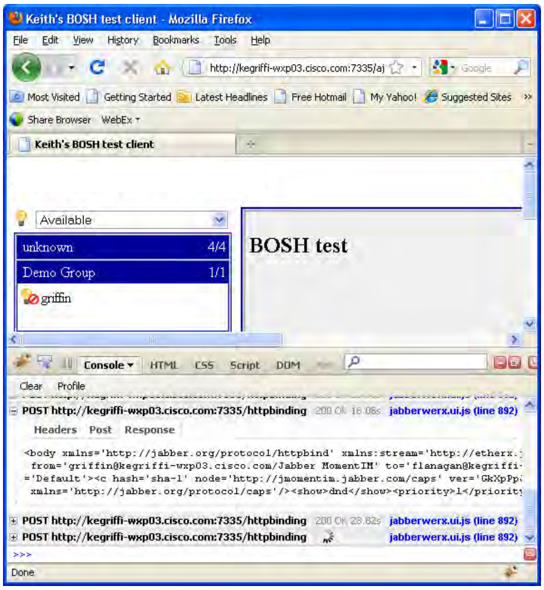


Fig. 6 Presence enabled client with BOSH Long Polling

It can be seen from the debug output in Fig. 6 that a presence state change to DND for user "griffin" has been received. Furthermore, it can be seen that the update was received in response to a HTTP POST request sent 16.08 seconds earlier. It can also be seen in the debug output that the following long poll expired without receiving an event before the 30 second timeout and the subsequent long poll is held open awaiting an asynchronous event from the server.

In both examples presence state changes were observed on browser-based clients and installed desktop clients based on an underlying protocol stack. In the case of continuous polling the installed client was SIP based and in the case of long-polling the installed client was XMPP based. What has been shown in the results for the sake of brevity is an example of a single state change. While intentionally not quantified by measurement, no difference could be observed to the human eye for presence state changes in the browser-based client. However it is expected that the use of a continuous polling technique would result in degraded performance and scalability of such a system when compared with a system using a long polling technique.

V. EVALUATION

Considering the nature of real-time communication events, in particular presence, it is clear that a mechanism is needed to allow browser-based communication clients to receive asynchronous real-time events in a manner similar to desktop clients which can use a dedicated protocol handler. Our goal is not to attempt to select the best web based asynchronous event mechanism for general purpose use as it is expected that research in this area will continue for quite some time. However it is necessary to evaluate available options that are capable of providing asynchronous events to browser-based communication applications.

All of the approaches outlined in this paper share the goal of enabling the delivery of asynchronous events to browser-based client applications. Continuous polling can be wasteful of network and server resources by polling continuously even when no data is available. Ajax is a useful technique when used to update a particular data set such as presence state while avoiding the need for a full page refresh. Long polling techniques such as the Bayeux Protocol and BOSH respect the fundamentals of HTTP such as only maintaining two connections with the server. Long polling also offers a more efficient use of network and server resources by handling asynchronous events over time and event controlled request-response pairs. It should be noted that for high asynchronous event rates the behavior of a given long polling technique approaches that of continuous polling as a high rate of asynchronous events result in an equivalent high rate of client driven requests. This is perhaps a useful reminder that while useful, long polling is itself a work-around technique as opposed to a true solution for the delivery of asynchronous events to browser-based applications. For the purposes of this research BOSH is noteworthy as not only is it a long polling technique that has been proven to work for instant messaging and presence based communication systems but it is also designed to be fully compliant with HTTP 1.0 and therefore suitable for use with constrained clients. While the long polling technique can clearly offer a suitable solution for asynchronous presence event delivery to browser-based applications the promise of bi-directional messaging via HTTP using the websocket technique to utilize the underlying TCP connection is promising. However, at time of writing the websocket standard is still being developed and it is too early to draw significant conclusions for its use with browser-based presence applications. It is suggested that, at the time of writing, long polling is the most suitable mechanism that can be used to deliver asynchronous real-time communication events to browser-based applications.

VI. CONCLUSION

We set out to identify a mechanism that would be suitable for the delivery of asynchronous real-time communication events, namely presence events to browser-based applications. After evaluating several techniques, including some recent and still under development, it is clear that it is certainly possible to deliver asynchronous presence events to browser-based applications. While basic continuous polling techniques can be used, they cannot scale. Web sockets offers the promise of a "more native" solution in time, however currently it is suggested that long polling techniques (the Bayeux protocol or BOSH) are a suitable mechanism for the delivery of asynchronous real-time events to browser-based applications.

VII. FUTURE WORK

While a suitable mechanism has been found for asynchronous real-time event delivery to browser-based applications this mechanism has been identified as a work around technique rather than a solution. An obvious area for future research would be the pursuit of a true solution for real bi-directional messaging between web client and server which could be used by real-time communication systems.

ACKNOWLEDGMENTS

With thanks to David Mathieson and Seamus Kerrigan for their tutorials on the implementation of JavaScript with Comet and XCP with BOSH respectively.

REFERENCES

[1] Jabber. The Power of Presence. 2008, http://www.jabber.com/CE/VideoVoiceCommunications.

[2] Rosenberg, J. A Presence Event Package for the Session Initiation Protocol (SIP). The Internet Society 2004, http://www.ietf.org/rfc/rfc3856.txt.

[3] Beltran, V. and J. Paradells, Middleware-based solution to offer mobile presence services, in Proceedings of the 1st international conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications. 2007, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering): Innsbruck, Austria.

[4] Berners-Lee, T., R. Fielding, and H. Frystyk. HTTP – Hypertext Transport Protocol (HTTP 1.0). RFC1945 1996 [cited; Available from: ftp://ftp.isi.edu/in-notes/rfc1945.txt.

[5] Fielding, R., et al. HTTP – Hypertext Transport Protocol (HTTP 1.1). RFC2616 1999, ftp://ftp.isi.edu/in-notes/rfc2616.txt

[6] Brown, K., et al. Web Services Polling (WS-Polling). 2005, http://www.w3.org/Submission/ws-polling/.

[7] Franklin, M. and S. Zdonik, Data in your face: push technology in perspective, in Proceedings of the 1998 ACM SIGMOD international conference on Management of data. 1998, ACM: Seattle, Washington, United States.

[8] W3C. XMLHttpRequest. W3C Working Draft 20 August 2009, http://www.w3.org/TR/XMLHttpRequest/.

[9] Russell, A., et al. Bayeux Protocol Draft 1.0. 2007,

[10] Dojo. Cometd Bayeux Ajax Push. 2009, http://alex.dojotoolkit.org/2006/03/comet-low-latency-data-for-the-browser/.

[11] Paterson, I., et al. XEP-0124: Bidirectional-streams Over Synchronous HTTP (BOSH). 2005, http://xmpp.org/extensions/xep-0124.html.

[12] Saint-Andre, P. XMPP Extension Protocols. 2001. XEP-0001. Available from: http://xmpp.org/extensions/xep-0001.html

[13] Paterson, I. and P. Saint-Andre. XEP-0206: XMPP Over BOSH. 2007, http://xmpp.org/extensions/xep-0206.html.

[14] Hickson, I. and D. Hyatt. A vocabulary and associated APIs for HTML and XHTML. 2009, http://www.w3.org/TR/html5/.

[15] Hickson, I. The Web Sockets API. 2009, http://dev.w3.org/html5/websockets/.

[16] Hickson, I. Server-Sent Events. 2009, http://dev.w3.org/html5/eventsource/.

[17] Hickson, I. The Web Socket protocol, 2009, http://tools.ietf.org/html/draft-hixie-thewebsocketprotocol-44.

[18] Cisco, Developer Guide for Cisco Unified Presence v7.0. 2008.

[19] Jabber. Jabber Extensible Communications Platform. 2008, http://www.jabber.com/CE/JabberXCP.

An Ontology-based Intrusion Detection for RFID Systems

M. Esposito[1,2], G. Della Vecchia[1]
[1] ICAR-CNR, Via Castellino 111, 80131 Naples, Italy
{massimo.esposito, gennaro.dellavecchia}@na.icar.cnr.it
[2] University of Naples "Parthenope", Centro Direzionale, Isola C4, 80143, Naples, Italy
massimo.esposito@uniparthenope.it

***Abstract*- In the last decade, RFID systems have gained increasing popularity, but they also pose many critical security concerns. The most challenging security threat in several RFID applications is tag cloning. In this paper, we do not focus on preventing tag cloning, but propose the application of the intrusion detection model to identify when tag cloning has occurred. In particular, we present an ontology-based misuse detection system (IDS) that integrates information coming from RFID middleware layer to detect tag cloning. Ontologies and rules are applied to formalize the declarative and procedural knowledge required to implement a "track & trace" technique, with the final aim of modelling two levels of detection against tag cloning. An inference engine is used to provide the proposed IDS with the advanced capacity of automatically reasoning about ontologies and rules in order to actually apply the formalized model of detection and infer when a tagged object is cloned or victim of a cloning attack.**

I. INTRODUCTION

In the last decade, RFID (Radio Frequency IDentification) systems have gained increasing attention, due to the growing number of investigations into different scenarios, varying from personnel tracking and localization to healthcare monitoring [1-3]. Nevertheless, there are still a vast number of problems that need to be solved before their massive deployment. As a matter of fact, an intense public concern with security issues has gone largely unaddressed and poses critical challenges to several RFID applications.

From the point of view of RFID technology, the most challenging security threat in several RFID applications is tag cloning. The research community addresses this threat primarily by applying variety of countermeasures, such as tag cryptography, authentication and access control.

The fundamental difficulties of such approaches revolve around the trade-off between tag cost, level of security, and hardware functionalities. As a matter of fact, RFID tags are typically deployed in numbers of several millions and the end-user companies have a strong financial incentive to minimize the tag cost and, thus, the features the tags provide [4].

As a result, it is extremely challenging to use cryptography for protecting low cost tags from cloning, due to their limited power, storage and processing resources.

Moreover, in order to supply cryptographic components with sufficient power, tags would need to be read from a shorter distance, which would degrade the read-rate of readers [5].

Although incremental improvements to the aforementioned trade-off, many reasons to assume that the existing solutions are not be proficient to completely protect tags from cloning can be pointed out.

Firstly, solutions instanced on the tag are intrinsically not secure, because, in an RFID system, tags are the weakest link in the whole chain due to their limited functional capabilities: an attacker, even with poor resources, can violate their security quite easily. Moreover, it is disputable whether it will be possible to produce a truly secure RFID tag, able to address all known vulnerabilities without improving the overall cost. Secondly, existing solutions mainly focus on preventing tag cloning, but a secure RFID system should go beyond prevention measures, by providing detection capabilities to identify when tag cloning has occurred [6].

In this work, we present an ontology-based intrusion detection system (IDS) that integrates information coming from RFID middleware layer to detect tag cloning. More specifically, we propose a misuse detection approach that implements a "track & trace" technique, essentially based on the reasoning that "when you know where a genuine tagged object is, a cloned ones can be detected". Such a technique has been implemented by applying ontologies in conjunction with rules, respectively to i) develop a formal description of a "track & trace" model and ii) formalize the declarative and procedural knowledge modelling two levels of defence against tag cloning. Ontologies and rules are utilized in conjunction with an inference engine to provide the IDS with the advanced capacity of automatically reasoning about the defined "track & trace" model, in order to infer if a tagged object is authentic, cloned or currently under attack.

The rest of the paper is organized as follows. Section II discusses related work. Section III describes our proposal of ontology-based intrusion detection for RFID systems. Section IV overviews our IDS architecture and outlines the implementation details. Finally, section V concludes the paper.

II. RELATED WORK

Despite the benefits of RFID, it is essentially an insecure technology. In particular, tag cloning is one of the most serious threats to the security of RFID systems. Tag cloning, i.e. when an attacker makes an exact copy of an RFID tag, enables an attacker to simply obtain the tag unique identifier, typically not kept secret on the majority of tags, so as to make

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_81,

their cloned tags indistinguishable from the originals. Once legitimate tag data has been obtained, attackers can reproduce their cloned tags on a wide scale.

In high cost RFID tags, where resources are not very restricted, several countermeasures have been devised to combat tag cloning, such as deactivation of tags, encryption, authentication and hash codes [7]. In [8], Juels has demonstrated some techniques for strengthening the resistance of EPC tags against cloning attacks, using PIN-based access to achieve challenge response authentication. In [9], Weis et al. have proposed a cryptographic approach to lock the tag without storing the access key, but only a hash of the key on the tag. The key is stored in a back-end server and can be found using the tag's meta-ID.

Avoine et al. [10] have proposed another hash-based RFID protocol that provides modified identifiers for improved privacy and that can be applied for authentication. In addition, hash-based RFID protocols for mutual authentication have been proposed in [11-12]. All these protocols rely on synchronized secrets residing on the tag and back-end server and require a one-way hash function from the tag.

In contrast, in low cost tags, due to their small size and constrained resources, complex cryptographic solutions like hash functions can not be executed. In [13] the authors mention scarcity of tag resources in low-cost RFID systems as a primary challenge in providing security and privacy mechanisms, and, also, in combating cloning. To address this issue, few lightweight authentication protocols that do not require cryptographic hash/keys in the tag have been proposed [14-15]. Another approach to combat cloning is the use of a Physical Unclonable Function (PUF) [16]. PUFs significantly increase physical security by generating volatile secrets that only exist in a digital form when a chip is powered on and running. Its main property is that it is easy to generate but hard to characterize.

All the cited solutions aim at preventing tag cloning. Additionally, tag cloning can be also detected by correlating information in the middleware layer. In such a sense, a very interesting approach has been proposed by Mirowski et al. [6]. More precisely, they have proposed an intrusion detection system for RFID systems, named Deckard, based on a statistical classifier and focused on the detection of change of tag ownership. This is one of the preliminary research to address the need for intrusion detection systems in RFID. Another remarkable approach, similar to Deckard in its intrusion detection architecture, has been proposed in [17]. Such an approach focuses beyond change in tag ownership and provides a more generic security framework to detect various RFID attacks.

The approaches outlined in [6, 17] are very remarkable for a several of ways. Firstly, they go beyond preventing tag cloning by actually determining whether the tag cloning has occurred. Secondly, they avoid the difficulties onboard the tags, by working with the middleware. As the reader and middleware components are typically accepted as the expensive components of RFID, such proposals are very practical [6].

Our approach closely resembles such an idea of applying the intrusion detection to identify tag cloning. But, differently, our research efforts have been primarily focused on developing an intelligent misuse detection system based on ontologies and rules, that, applied in conjunction with an inference engine, provide the advanced capacity of automatically reasoning to detect tag cloning. To the best of our knowledge, the RFID security literature has not yet addressed applications of ontologies and rules to implement intrusion detection in RFID systems, neither system-oriented researches appear to have been developed in that direction.

III. THE ONTOLOGY-BASED INTRUSION DETECTION APPROACH

A. *Misuse Detection*

Traditionally, there are two basic approaches to intrusion detection: anomaly detection and misuse detection. In anomaly detection, the goal is to define and characterize legitimate behaviours of the users, and, then, detect anomalous behaviours by quantifying deviations from the former by means of established metrics. Differently, misuse detection involves the comparison of a user's activities with known behaviours of attackers, formalized in different models.

The approach proposed in such a work relies on the application of a misuse detection based on ontologies and rules to identify tag cloning in RFID systems. Such a choice is due to a several of issues.

Firstly, a misuse detection based on ontologies and rules enables to have a clear understanding of what constitutes a scenario of tag cloning, and, as a result, it is characterized by a low false positive rate. Secondly, it is more intuitive, particularly in the case when users need to formalize and understand their knowledge about specific scenarios. This involves that a misuse detection system based on ontologies and rules knows how tag cloning should manifest itself. This leads to a simple and efficient processing of the audit data. The obvious disadvantages of such a model rely on the time-consuming task of specifying ontologies and rules and the scarce ability to identify novel situations of tag cloning.

Nevertheless, we believe that, in the context of our scenario, such issues do not constitute drawbacks and the proposed misuse detection can be proficiently applied, due to the very limited amount of available data that, thus, significantly facilitates the task of specifying and keeping ontologies and rules up to date.

B. *The "track & trace" technique*

The misuse detection has been actually realized by applying the "track & trace" technique. This technique relies on the reasoning that "when you know where the genuine tagged object is, the fake/cloned ones can be detected".

In detail, "track & trace" refers to generate and store inherently dynamic profiles of RFID tagged object's movements through the supply chain. Such profiles are built exploiting the set of location events related to one tagged object that can be retrieved from a tracing system, such as the EPC network [18]. In addition to such dynamic profiles, tagged objects are also characterized by static profiles, i.e. fixed paths composed of point of interests (POI) through the supply chain. Moreover, POIs specify normal usage conditions for tagged objects.

According to such profiles, it is possible to apply two levels of detection against tag cloning. More precisely, a first level of defence refers to detect all the tagged objects that have an invalid tag serial number. The second level of defence is to detect whether a tagged object introduced into the supply chain is cloned or currently under a cloning attack.

The first situation, i.e. a cloned tag is detected, occurs when:

- a genuine tagged object is in multiple locations at the same time, that means it is expected to be at a particular point, and, at the same time, a tagged object with the same ID is scanned at a different point in the supply chain;
- a genuine tagged object moves in the supply chain too fast, that means it is scanned at a particular point, and, then, scanned again at a different point after a too short time interval.

The second situation, i.e. a genuine tag is recognized as victim of a cloning attack, occurs when:

- a genuine tagged object moves in the supply chain in an abnormal way, that means it does not move according to the fixed path defined in the static profile, but goes, for example, from the stocking room of a clothing shop to its bathroom;
- a genuine tagged object is utilized in the supply chain in an abnormal way, that means it is scanned at a particular point too many times in a fixed time window with respect to the expected usage condition.

In order to formalize both the static and dynamic profiles to be associated to each tagged object, we have devised a “track & trace” model, based on the ideas of physical and semantic locations. These notions are not new in literature [2], but we have partially re-elaborated them.

Generally, a physical location specifies a precise region identified by means of the proximity to well-known points. Differently, a semantic location specifies the meaning of a location and usually covers more physical locations. More specifically, physical locations are the areas covered by RFID readers, whereas a semantic location can be a country, a city, a building, a room inside a building, a railway station and so on. According to the typical use of RFID technology in indoor situations, we have defined semantic locations for a generic indoor environment, i.e. building, room, corridor, and so on. Each of these semantic locations is related to the city and country where it is placed in.

The static profile to be associated to each tagged object represents a fixed path of POIs, structured in terms of semantic locations to be transited through the supply chain. An example of fixed path is shown in Fig.1. Such a path is composed of four notable POI: the first one is located in France, the second one in Switzerland, the third and the last ones in Italy. Each of these POIs has to be described in detail in terms of country, city, address, building and room inside the building, such as reported, as an example, in Fig.1.

Fig. 1. An example of RFID tagged object path

Such a fine granularity of formalization is required because it is worth specifying, for example, which room inside a clothing shop at “Piazza del Plebiscito” in Naples the tagged object is placed in. As a matter of fact, if the room inside the clothing shop is different from the expected one, it is possible that the tagged object is currently victim of a cloning attack.

In contrast, the dynamic profile of a tagged object represents an enlarging path of POIs, structured in terms of physical locations transited through the supply chain. Physical locations can be inferred by means of the audit records outcome by RFID read/write operations. As a matter of fact, a typical audit record can be logically structured as follows: *<tag number, reader number, RFID operation, timestamp>*. According to such data, it is possible to infer the physical location of an RFID tagged object for proximity to the reader that has performed the operation.

In order to determine which semantic location corresponds to a physical one, semantic locations have been subdivided in many physical locations, that represent the minimal area in which a tag can be localized, as shown in Fig.2.

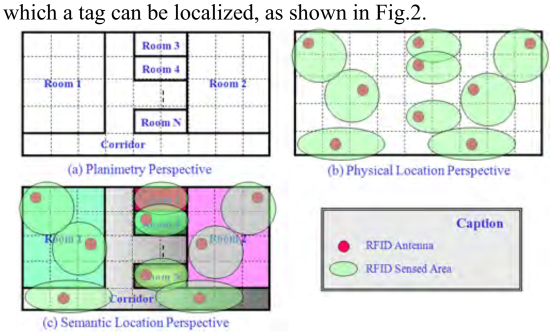


Fig. 2. Representation of the physical and semantic location perspectives

As a result, according to such a formalization and to the physical locations obtained from the audit records, it is easily to identify, for a tagged object, the corresponding semantic location, and, thus, verify the consistency of its movements with respect to its static profile. Moreover, for a tagged object, an expected usage condition is associated to each POI of its static profile: if, for example, in the warehouse of the clothing shop at “Piazza del Plebiscito”in Naples, the number of readings related to an RFID tagged object exceeds a threshold indicating a normal behaviour, it is possible that the tagged object is currently victim of fraudulent readings.

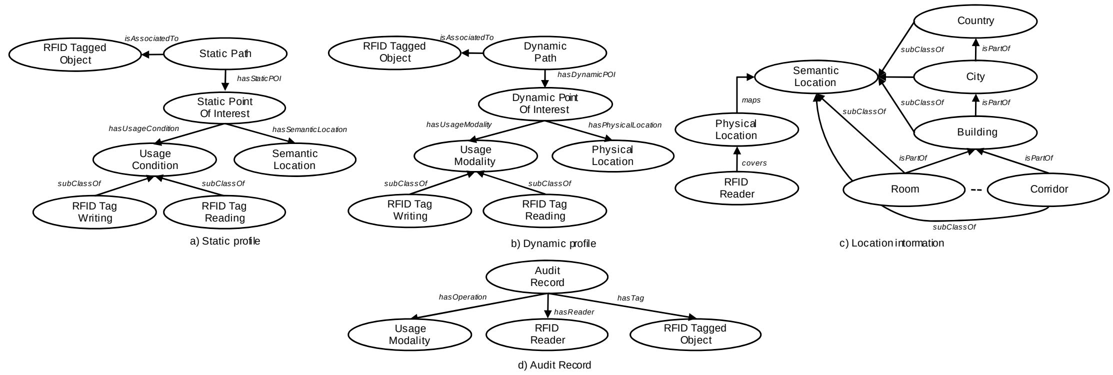


Fig. 3. "Track & Trace" Ontology

C. *Ontologies and rules for the "track & trace" model*

The "track & trace" model specifies information regarding static and dynamic profiles, locations and audit records in a unique and sharable manner. Nevertheless, this information must be represented in an unambiguous and well-defined formalism and expressed in a machine-readable format in order to facilitate the application of reasoning techniques to compare static and dynamic profiles of tagged objects.

As a result, we have defined an ontology to formally describe the "track & trace" model, in terms of relevant concepts, properties and interrelations. More precisely, Fig.3 presents a high level view of the defined ontology, named "Track and Trace" Ontology, and formalized in standard OWL (Web Ontology Language) [19].This ontology is subdivided in four logical parts for sake of clarity.

The first part of the ontology is intended to model a static profile of a tagged object. In Fig. 3 (a), the main concepts and properties are outlined. A complete and detailed list of properties of each concept and sub-concept (denoted by ellipses) is not shown in figure, because it would make the representation unwieldy, but it is reported in Tab. 1.

Role	Domain	Range	Inverse	Trans.
hasStaticPOI	StaticPath	StaticPoint OfInterest	isStatic POIOf	no
hasStaticPathID	StaticPath	*Datatype*: String	-	-
isAssociatedTo	StaticPath	RFIDTaggedObject	has StaticPath	no
hasUsageCondition	StaticPointOf Interest	UsageCondition	isUsage ConditionOf	no
hasSemanticLocation	StaticPointOf Interest	SemanticLocation	isSemantic LocationOf	no
hasStaticPOIID	StaticPointOf Interest	*Datatype*: String	-	-
hasTagID	RFIDTaggedObject	*Datatype*: String	-	-
hasTagData	RFIDTaggedObject	*Datatype*: String	-	-
hasSemantic LocationName	SemanticLocation	*Datatype*: String	-	-
hasConditionName	UsageCondition	*Datatype*: String	-	-
hasMaxValue	UsageCondition	*Datatype*: Integer	-	-

Tab. 1 Properties for modelling a static profile

Each property is defined in terms of domain (the set of possible subject concepts) and range (the set of possible object concepts or datatypes). Besides, the inverse property is reported, if it is admissible, and the transitiveness is specified, if it exists. It is worth noting that each sub-concept inherits super-concept properties and adds new specialized ones.

The second part of the ontology models a dynamic profile. The main concepts and properties are outlined in Fig. 3 (b), whereas a more detailed list of properties is reported in Tab. 2.

Role	Domain	Range	Inverse	Trans.
hasDynamicPOI	DynamicPath	DynamicPoint OfInterest	isDynamic POIOf	no
hasDynamicPathID	DynamicPath	*Datatype*: String	-	-
isAssociatedTo	DynamicPath	RFIDTaggedObject	has DynamicPath	no
hasUsageModality	DynamicPoint OfInterest	UsageModality	isUsage ModalityOf	no
hasPhysicalLocation	DynamicPoint OfInterest	PhysicalLocation	isPhysical LocationOf	no
hasOperation Counter	UsageModality	*Datatype*: Integer	-	-
hasTimestamp	UsageModality	*Datatype*: Time	-	-
hasTagID	RFIDTaggedObject	*Datatype*: String	-	-
hasTagData	RFIDTaggedObject	*Datatype*: String	-	-
hasValidID	RFIDTaggedObject	*Datatype*: Boolean	-	-
isCloned	RFIDTaggedObject	*Datatype*: Boolean	-	-
isUnderAttack	RFIDTaggedObject	*Datatype*: Boolean	-	-
hasPhysical LocationID	PhysicalLocation	*Datatype*: String	-	-

Tab. 2 Properties for modelling a dynamic profile

The third part of the ontology is devised to specify location information by means of concepts, as outlined in Fig. 3 (c), and properties, as diffusely reported in Tab. 3.

Role	Domain	Range	Inverse	Trans.
isPartOf	SemanticLocation	SemanticLocation	hasPart	yes
hasLocationName	SemanticLocation	*Datatype*: String	-	-
hasPhysical LocationID	PhysicalLocation	*Datatype*: String	-	-
maps	PhysicalLocation	SemanticLocation	isMappedWith	-
covers	RFIDReader	PhysicalLocation	isCoveredBy	no
hasReaderID	RFIDReader	*Datatype*: String	-	-
hasAddress	Building	*Datatype*: String	-	-
hasFloors	Building	*Datatype*: Integer	-	-
hasRooms	Building	*Datatype*: Integer	-	-
isOnFloor	Room	*Datatype*: Integer	-	-
isBorderingOn	Room	Room	-	-

Tab. 3 Properties for modelling location information

Finally. the last part of the ontology specifies information contained in an audit record by means of concepts, as outlined in Fig. 3 (d), and properties, as diffusely reported in Tab. 4.

Role	Domain	Range	Inverse	Trans.
hasOperation	AuditRecord	RFIDOperation	isOperationOf	no
hasTag	AuditRecord	RFIDTaggedObject	isTagOf	no
hasReader	AuditRecord	RFIDReader	isReaderOf	no
hasTimestamp	AuditRecord	*Datatype*: Time	-	-

Tab. 4 Properties for modelling an audit record

Moreover, according to such an ontology, we have formalized, in a set of SWRL (Semantic Web Rule Language) [20] rules, the declarative and procedural knowledge modelling the two levels of detection previously described. For sake of brevity and clarity, we have grouped the implemented rules in logical sets and reported their description only in natural language:

- *Analysis and processing of data contained in audit records*. Depending on data contained in an audit record, diverse sets of rules have been formalized for:
  - identifying the dynamic profile to be updated with new information;
  - determining the physical location of the sensed tag for proximity to the reader that has performed the operation;
  - adding a new dynamic POI in the case when the detected physical location belong to no POI. Such a POI will be also characterized by the usage modality and the relative timestamp, both extracted by the audit record;
  - updating an existing dynamic POI in the case when the detected physical location belong to an existing POI, but a new operation has been performed. The existing POI will be characterized by the new usage modality and the relative timestamp, both extracted by the audit record;
  - updating the usage modality of an existing dynamic POI in the case when the detected physical location belong to an existing POI, and the operation has been already performed in the past. The existing POI will be updated with the timestamp and the occurrences of the last performed operation.
- *Execution of the first level of detection*. A set of rules have been formalized for detecting all the tagged objects that have an invalid tag serial number.
- *Execution of the second level of detection.* Depending on data stored in both the static and dynamic profiles, sets of rules have been formalized for:
  - determining the correspondence existing between semantic and physical locations;
  - inferring that a tagged object is cloned in the case when it is detected simultaneously in at least two different and not bordering dynamic POIs in a too short time window;
  - inferring that a tagged object is cloned in the case when a dynamic POI is not expected in the static profile and it is not bordering with static POIs;
  - inferring that a tagged object is under attack in the case when a dynamic POI is not expected in the static profile, but it is bordering with static POIs;
  - inferring that a tagged object is under attack in the case when the usage modality of a dynamic POI is not in accordance with the usage condition of the corresponding static POI, in terms of number of operations allowable in a pre-defined time window.

An inference engine is utilized in conjunction with such formalized ontologies and rules to automatically reason about static and dynamic profiles of tagged objects and, thus, to implement the two reported levels of detection against tag cloning.

IV. THE ONTOLOGY-BASED IDS ARCHITECTURE

The proposed IDS has been devised to be transparently and seamlessly integrated into a given RFID system by means of being positioned at the middleware level in order to effectively shield back-end systems from cloning attacks. Thus, even if a cloned tag enters the reader field, its data will not propagate beyond the IDS, the backend will never be aware of its presence and normal operations will continue as if nothing happened. The architecture of the proposed IDS consists of the set of components described below:

- *Target System*. It is a typical RFID system capable of producing data that summarizes activity of tags in its coverage area. It is responsible for recording the outcome of a RFID read/write operation into an audit record.
- *Knowledge Base*. It contains the formalization of the domain knowledge regarding the possible scenarios of tag cloning. It consists of three entities: i) a Terminological Box, populated with the previously described OWL ontologies, ii) the Rule Box, populated with the previously described SWRL rules, and iii) the Assertional Box, populated with the audit records, appropriately formalized in RDF (Resource Description Framework) [21].
- *Auditing Module.* It is the component responsible for collecting and logging the audit records received from the target system into the Assertional Box of the Knowledge Base.
- *Decision Module*. It is the smart component that implements the "track & trace" technique by applying ontologies and rules, as described in the previous section. It relies on the logic inference engine presented in [22], able to integrate and reason about ontologies and rules to perform inferential reasoning on the domain knowledge stored into the Knowledge Base.

With such components in mind, the proposed IDS operates in the following manner. Firstly, the Target System records the details of a new operation into an audit record. This record is, then, stored by the Audit Module inside the Assertional Box of the Knowledge Base. Secondly, according to the data contained in the new audit record, the Decision Module updates the dynamic profile of the appropriate tag and activates the inference engine in order to compare the appropriate static and dynamic profiles stored into the

Knowledge Base, and, thus, infer if the sensed tag is authentic, cloned or currently under attack.

An experimental prototype for testing purpose has been implemented at the Institute for High Performance Computing and Networking (ICAR) of Italian National Research Council (CNR). All the described components have been implemented as fully portable Java entities, with the exception of the Audit Module, which also integrate, as native code, Microsoft Visual C++ 6.0 control drivers provided by the RFID technology manufacturers. Specifically, the Target System consists of RFID readers belonging to FEIG's Passive RFID Long Range Reader series, model ISC.LRU2000-A. These readers are able to interact with EPC Class 1 Gen 2 transponders for operating frequencies in the UHF range 865-870 MHz and reading range of up to 10 meters (approx. 33 feet). FEIG's UHF antennas model ID ISC.ANT.U250/250-EU have been installed, too. These antennas are circular polarized for operating frequencies in the UHF range 865-870 MHz.

Preliminary tests performed on the experimental prototype aimed at detecting synthesized cloning attacks gave a proof of the feasibility of our approach, suggesting that an actual deployment of the proposed IDS can effectively support the detection of cloned tags in RFID systems.

V. CONCLUSION

Detection may be seen as the first step in defending against tag cloning and preventing RFID-enabled crime from occurring. In this perspective, the present paper has described a misuse detection approach, based on the "track & trace" technique, for RFID systems. Ontologies and rules have been used in order to implement such a technique, by enabling the formalization of the declarative and procedural knowledge required to model two levels of defence against tag cloning. Moreover, an inference engine has been used to provide the proposed IDS with the advanced capacity of automatically reasoning about ontologies and rules in order to actually apply the formalized model of detection, and infer when a tagged object is cloned or victim of a cloning attack.

As concluding remarks, it is relevant to observe that the approach described in this paper represents only a preliminary step of an ongoing research, because, at the moment, an experimental prototype has been tested only over proof cloning attacks. Next step of our research will be to minutely investigate the feasibility of the approach in real cases, in terms of performance evaluation and experimental assessment.

REFERENCES

[1] M. Ciampi, A. Coronato, G. De Pietro, M. Esposito, "A Location Service for Pervasive Grids", in Proceedings of CISSE 2005, Book "Advances in Systems, Computing Sciences and Software Engineering", Springer, 2006, pp. 119-123, DOI 10.1007/1-4020-5263-4_19.

[2] A. Coronato, M. Esposito, G. De Pietro, "A Multimodal Semantic Location Service for Intelligent Environments: An Application for Smart Hospitals", Journal of Personal and Ubiquitous Computing (JPUC), Springer-Verlag, 2009, vol. 13, no 7, pp. 527-538, DOI 10.1007/s00779-009-0223-x.

[3] G. Della Vecchia, M. Esposito, "A Pervasive System for Nuclear Medicine Departments", Journal of Wireless Personal Communications (WPC), Springer, 2009, DOI 10.1007/s11277-009-9789-x.

[4] M. Lehtonen, D. Ostojic, A. Ilic, and F. Michahelles, "Securing RFID Systems by Detecting Tag Cloning", in Proceedings of the 7th International Conference on Pervasive Computing, Lecture Notes In Computer Science, Springer-Verlag, 2009, vol. 5538,pp. 291-308. DOI 10.1007/978-3-642-01516-8_20.

[5] D.C. Ranasinghe, D.W. Engels, and P.H. Cole, "Low-Cost RFID Systems: Confronting Security and Privacy", in Auto-ID Labs Research Workshop, Zurich, Switzerland, 2005.

[6] L. Mirowski, and J. Hartnett, "Deckard: A system to detect change of RFID tag ownership", International Journal of Computer Science and Network Security, 2007, vol. 7, pp. 89-98, ISSN 1738-7906 (2007).

[7] T. Karygiannis, B. Eydt, G. Barber, L.Bunn, and T. Phillips, "Guidelines for securing radio frequency identification (RFID) systems," April 2007, NIST Special Publication 800-98.

[8] A. Juels, "Strengthening EPC tags against cloning", in Proceedings of WiSe '05 (Cologne, Germany, September 2005), ACM, New York, NY, 2005, pp. 67-76, DOI 10.1145/1080793.1080805.

[9] S. Weis, S. Sarma, R. Rivest, and D. Engels, "Security and Privacy Aspects of Low-cost Radio Frequency Identification Systems". in Proceedings of SPC 2003, Lecture Notes in Computer Science, Springer-Verlag, 2004, vol. 2802, pp. 50-59, DOI 10.1007/b95124.

[10] G. Avoine, and P. Oechslin, "A scalable and provably secure hash based RFID protocol", in Proceedings of PerSec 2005, IEEE Computer Society Press, 2005, pp. 110–114, DOI 10.1109/PERCOMW.2005.12.

[11] E.Y. Choi, S.M. Lee, and D.H. Lee, "Efficient RFID authentication protocol for ubiquitous computing environment", in Proceedings of SecUbiq 2005, Lecture Notes in Computer Science, Springer-Verlag, 2005, vol. 3823, pp. 945-954, DOI 10.1007/11596042.

[12] S. Lee, T. Asano, and K. Kim, "RFID Mutual Authentication Scheme based on Synchronized Secret Information", in Proceedings of In Symposium on Cryptography and Information Security, Hiroshima, Japan, January 2006.

[13] S. Sarma, S. Weis, and D. Engels, "Radio-frequency identification: Security risks and challenges", RSA Laboratories Cryptobytes, 2003, vol. 6, no 1, pp. 2-9.

[14] S. Karthikeyan, and M. Nesterenko, "RFID security without extensive cryptography", in Proceedings of SASN'05, ACM, New York, NY, 2005, pp. 63-67, DOI 10.1145/1102219.1102229.

[15] H.-Y. Chien, "SASI: A new ultralightweight RFID authentication protocol providing strong authentication and strong integrity", IEEE Transactions on Dependable and Secure Computing, 2007, vol. 4, no 4, pp. 337-340, DOI 10.1109/TDSC.2007.70226.

[16] S. Devadas, E. Suh, S. Paral, R. Sowell, T. Ziola, and V. Khandelwal, "Design and implementation of PUFbased "unclonable" RFID ICs for anti-counterfeiting and security applications", in Proceedings of the 2008 IEEE international conference on RFID, IEEE Computer Society Press, 2008, pp. 58–64, DOI 10.1109/RFID.2008.4519377.

[17] G. Thamilarasu, R. Sridhar, "Intrusion detection in RFID systems", in Proceedings of MILCOM 2008, IEEE Computer Society Press, 2008, pp. 16-19, DOI 10.1109/MILCOM.2008.4753218.

[18] EPCglobal Inc. "EPCglobal Architecture Framework Version 1.3", http://www.epcglobalinc.org/standards/architecture/architecture_1_3-framework-20090319.pdf.

[19] Web Ontology Language, available at www.w3.org/TR/owl-features/ .

[20] Semantic Web Rule Language, available at www.w3.org/Submission/SWRL/ .

[21] Resource Description Framework, available at www.w3.org/RDF/ .

[22] M.Esposito, "An Ontological and Non-monotonic Rule-based Approach to Label Medical Images", in Proceedings of SITIS 2007, IEEE Computer Society Press, 2007, pp. 603-611, DOI 10.1109/SITIS.2007.77.

A New Algorithm and its Implementations for RC Filters

GUOLIANG ZENG
Phone: (480) 727-1905; Fax: (480) 727-1773; E-mail: gzeng@asu.edu
Motorola DSP Laboratory
Department Engineering
Arizona State University Polytech
7231 Sonoran Arroyo mall, Santan Hall, Mesa, AZ 85212

ABSTRACT: In digital communication, it has been an important topic to implement an RC (Raised Cosine) filter digitally. Based on the characteristics of the digital communication, a new symbol-based algorithm is proposed and implemented with MATLAB. It is also realized on Motorola's DSP56303 in assembly language. The new algorithm features the minimum data buffer and execution time, hence the minimum output time delay. Therefore, it is very practically useful for high-speed data transmission. This new method can be applied to any convolution and/or correlation of two digital signals with different sampling periods. It should be applied in all up- and down-sampling signal processing systems.

KEYWORDS: RC filter, Convolution, Correlation, Digital communication.

INTRODUCTION: In a digital communication system, information bits are carried by symbol pulses. Due to the finite bandwidth of a practical channel, a time-limited pulse, like square pulse will be distorted by the channel, causing Inter-Symbol-Interference (ISI). This is because a time-limited pulse has an infinite bandwidth of frequency. To achieve zero ISI in a practical system, each symbol pulse received must be a band-limited Nyquist pulse. The widely used band-limited Nyquist pulse is the family called Raised Cosine (RC) pulse.

A raised cosine (RC) pulse is given by the following equation [1]:

$$rc(t)=\frac{\sin(\pi t/T_s)}{\pi t/T_s}\frac{\cos(\pi\alpha t/T_s)}{1-(2\alpha t/T_s)^2}=\sin c(t/T_s)\frac{\cos(\pi\alpha t/T_s)}{1-(2\alpha t/T_s)^2},\ -\infty<t<\infty \tag{1}$$

Here, $0 \le \alpha \le 1$ is called roll-off factor. The T_s is the symbol period. The spectrum of the RC pulse is the Fourier transform of equation (1):

$$RC(\omega)=\begin{cases} T_s,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, |\omega|\le\pi(1-\alpha)/T_s \\ \frac{T_s}{2}\{1+\cos[\frac{1}{2\alpha}(T_s|\omega|-\pi(1-\alpha))]\}= \\ T_s\cos^2[\frac{1}{4\alpha}(T_s|\omega|-\pi(1-\alpha))],,,,,,,,,,,,,, \frac{\pi(1-\alpha)}{T_s}<|\omega|\le\frac{\pi(1+\alpha)}{T_s} \\ 0,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, |\omega|>\pi(1+\alpha)/T_s \end{cases} \tag{2}$$

Therefore the frequency bandwidth of an RC pulse is finite. It can pass some practical channels without distortion.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_82,

A filter is called an RC filter if its impulse response $h(t)$ is an RC pulse: $h(t) = rc(t)$. Clearly, an RC filter is a low-pass non-causal filter.

A well equalized (flat) digital transmission system can be modeled as an RC filter with the following input:

$$x(t) = \sum_{k=-\infty}^{\infty} d[k]\delta(t - kT_s), \quad -\infty < t < \infty \tag{3}$$

Here, $k = 0, \pm 1, \pm 2, \cdots$ is the index for symbols. T_s is the symbol period. And, $d[k]$ is the information data sequence. It consists of 1s and 0s for binary unipolar signaling, or 1s and –1s for binary bipolar signaling. The output is the convolution of input and impulse response:

$$y(t) = x(t) * h(t) = \sum_{k=-\infty}^{\infty} d[k]h(t - kT_s) \tag{4}$$

Note that the duration of $h(t)$ is from -∞ to ∞ as shown in (1). For a practical digital filter, it must be truncated, shifted, and sampled. Assume that $h(t)$ is truncated from $-DT_s$ to DT_s, for some integer D. And N samples are taken in each symbol period T_s. That is the sampling period is T_s/N. Therefore the impulse response sequence has $2DN$ samples:

$$h[n] := rc(n\tfrac{T_s}{N}), \quad n = -DN, -DN + 1, \cdots, DN - 1. \tag{5}$$

To get the output by convolution, the input $x(t)$ must be sampled with the same sampling period T_s/N:

$$x[n] := x(n\tfrac{T_s}{N}), \quad n = 0, \pm 1, \pm 2, \cdots. \tag{6}$$

The output sequence is the convolution of input sequence and impulse response sequence:

$$y[n] = x[n] * h(n) = \sum_{j=-DN}^{DN-1} h[j]x[n-j], \quad n = 0, \pm 1, \pm 2, \cdots \tag{7}$$

CURRENT ALGORITHM: Consider the input sequence $x[n]$ obtained by (6). From equation (3), every N elements of $x[n]$, there is only one nonzero element, which is the information data $d[k]$. Other N – 1 elements are zeros. In fact, the sequence of $d[k]$ is the only physical input. In practical implementations, for example, in the function "`rcosflt`" of MATLAB [2], $x[n]$ is formed by "upsampling" which pads N – 1 zeros for each $d[k]$. Since "`rcosflt`" uses built-in functions, it is rewritten for comparison as follows:

```
function [yo,to] = rcosflt1(x,
Fd, Fs, R, Delay)
%x = the input symbol sequence.
%The symbol rate is Fd(sym/sec).
%The sample rate is Fs(sam/sec).
%Fs must be an integer multiple.
%of Fd. R = the roll-off factor,
% R must be in the range [0, 1].
%DELAY is time delay.
%Y = the output sample sequence.
%
N=Fs/Fd;
% Design filter.
th=linspace(-Delay, Delay,
            2*Delay*N+1);
for i=1:length(th)
    if th(i)==0.0
      yf(i)=1.0;
    elseif abs(th(i)) == 1/
                 (2*R*Fd)
      yf(i)=(sin(pi/(2*R)))*
            (R/2);
    else
      yf(i)=((sin(pi*th(i)*Fd))
         *(cos(pi*R*th(i)*Fd)))
         / ((pi*th(i)*Fd).*(1-
```

```
          (2*R*th(i)*Fd).^2));
    end
end
% Upsample input x from Fd to Fs
%by padding N-1 0s per symbol.
lx=length(x);
xx = zeros(1,(lx+Delay*2)* N);
     for i = 1:lx
        xx((i-1)*N + 1) = x(i);
     end;
% Convolve xx and yf for output.
lxx=length(xx); lyf=length(yf);
lo=lxx+lyf-1;
if lyf > lxx
    tmpt=yf; yf=xx; xx=tmpt;
end

yo=[];
for n=0:lo-1
    temp=0.0;
    if n<lyf
        for m=0:n
            temp=temp+xx(m+1) *
                  yf(n-m+1);
        end
    elseif n<lxx
        for m=(n-lyf+1):n
            temp=temp+xx(m+1)*
                  yf(n-m+1);
        end
    else
        for m=(n-yf+1):(lxx-1)
      temp=temp+xx(m+1)*
            yf(n-m+1);
        end
    end
    yo=[yo temp];
end
to = [0:lo-1]/Fs;
```

The above current algorithm is sample-based, that is it processes signal sample by sample. This requires a lot of zeros padded in $x[n]$. They waste a lot of memory space and calculation time, causing bigger delay for the output. This limits the speed of data transmission. To overcome this shortcoming, a new algorithm is proposed as follows.

NEW ALGORITHM: Let us rewrite equation (5) as a $2D{\times}N$ matrix h:

$$h = (h_{ki})_{2D\times N}\text{, with } h_{ki} := rc((k+\tfrac{i}{N})T_s),$$

$$k = -D, -D+1, \cdots, D-1;\ i = 0, 1, \cdots, N-1 \quad (8a)$$

Here, integer k is the symbol index, integer i is the sample index in each symbol period. Therefore,

$$h = \begin{pmatrix} h_{-D0}, h_{-D1}, \cdots\cdots\cdots, h_{-D(N-1)} \\ h_{(-D+1)0}, h_{(-D+1)1}, \cdots, h_{(-D+1)(N-1)} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ h_{(D-1)0}, h_{(D-1)1}, \cdots\cdots, h_{(D-1)(N-1)} \end{pmatrix}_{2D\times N} \quad (8b)$$

At time $t = kT_s$, let the data vector be

$$d=(d[k+D],d[k+D-1],\ldots,d[k-D+1])_{1\times 2D}$$
$$k = 0,\pm 1,\pm 2,\cdots \quad (9)$$

Hence, the output is the sequence of the following vectors:

$$y_k = (d*h)_{1\times N},\ k = 0,\pm 1,\pm 2,\cdots \quad (10)$$

Here, y_k is a vector consisting of N elements. It is not difficult to check that equation (10) gives the same out put as $y[n]$ in (7). Equations (8), (9) and (10) form a new, symbol-based algorithm.

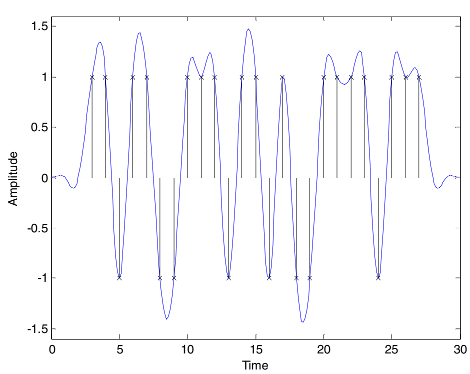
An algorithm specifies the arithmetic operations to be performed but does not specify how that arithmetic is to be implemented. Therefore, the implements of an algorithm are also important especially for engineers. Now, we implement the new algorithm with MATLAB:

```
function [yo,to] = rcosflt2(x,
Fd, Fs, R, Delay)
%x = the input symbol sequence.
%The symbol rate is Fd(sym/sec).
%The sample rate is Fs(sam/sec).
%Fs must be an integer multiple.
%of Fd. R = the roll-off factor,
% R must be in the range [0, 1].
%DELAY is time delay.
%Y = the output sample sequence.
%
N=Fs/Fd;
```

```
%Design impulse response matrix:
for k=1:2*Delay
    for i=1:N
        th=((k-Delay-1)+
            (i-1)/N)/Fd;
        if th==0.0
            h(k,i)=1.0;
        elseif abs(th)==1/
                  (2*R*Fd)
            h(k,i)=(sin(pi/
                  (2*R)))*(R/2);
        else
            h(k,i)=((sin(pi*th*
            Fd)).*(cos(pi*R*th*
            Fd)))./((pi*th*Fd).*
            (1-(2*R*th*Fd).^2));
        end
    end
end
y=[x' zeros(1,2*Delay)];
d=zeros(1,2*Delay);
yo=[];
for j=1:length(y)
    d=circshift(d,[0 1]);
    d(1)=y(j);
    yj=d*h;
    yo=[yo yj];
end
to = [0:length(yo)-1]/Fs;
```

Obviously, the new algorithm is symbol-based, that is it processes signal symbol by symbol. In contrast with the sample-based one, it only requires about one *N*th memory space and processing time (number of multiplications). Usually, N = 8 or 16.

The following plot shows the same result (output) of the above two algorithms with `Fd=1, Fs=8, R=0.5, Delay=3`.



Furthermore, since the decisions in digital communication are made symbol by symbol, the new algorithm can easily be realized in hardware for real-time processing as shown below.

HARDWARE REALIZATION: The architecture and addressing modes of a DSP chip are designed for vector (parallel) processing. Hence, the proposed new algorithm can be implemented with the DSP chips efficiently. The following assembly language program is tested with Motorola's DSP56303 in a real-time processing system [3].

```
;M = 2D = # of symbols for the
;impulse response h(t).
;N=# of samples in each symbol.
;x0 = input symbol.
;a = output sample.
;x:$ffef=SSI RX/TX data registr.
;Xbase = base location of filter
;states d[k] in X memory.
;Ybase=base location of impulse
;response h[k,i] in Y memory.

;Assume M=4, N=6:
      M     equ   4
      N     equ   6
;X memmmory:
            Org   x:Xbase
      List1 dc    d[-2], d[1],
                  d[0], d[-1]
      ;new data d[2] in x:$ffef
;Y memmmory:
            Org   y:Xbase
      List2 dc    h_{-2,0},h_{-2,1},h_{-2,2},
                  h_{-2,3},h_{-2,4},h_{-2,5},
                  h_{-1,0},h_{-1,1},h_{-1,2},
                  h_{-1,3},h_{-1,4},h_{-1,5},
                  h_{0,0}, h_{0,1}, h_{0,2},
                  h_{0,3}, h_{0,4}, h_{0,5},
                  h_{1,0}, h_{1,1}, h_{1,2},
                  h_{1,3}, h_{1,4}, h_{1,5}

      move   #Xbase,r0
      ;point to states d[k].
      move   #M-1,m0     ;mod(M).
```

```
      move   #Ybase,r4
      ;point to impulse rspns.
      move   #MN-1,m4
      ;mod(MN).
      move   #N,n4
      ;offset(N).

loop
      movep x:$ffef, x0
      ;read A/D.
      do     #N,doend
      clr    a      x0,x:(r0)+
             y:(r4)+n4,y0
      rep    #M-1
      mac    x0,y0,a
             x:(r0)+,x0
             y:(r4)+n4,y0
      macr   x0,y0,a     x(r0),x0
             (r4)+
doend movep a,x:$ffef
      ;write D/A
      operation   __, __, _
             (r0)-   (r4)-n4
      jmp    loop
```

CONCLUSION: This paper proposes a new, symbol-based algorithm and its implementations in MATLAB and on Motorola's DSP56303 for RC filters. Contrasting the traditional sample-based algorithm, the amount of reduction in memory space and execution time is directly proportional to *N*. Generally, *N* is greater than or equals 8. Therefore, the improvement is very notable.

REFERENCES:

[1] A. Bateman, *Digital Communications*, Addison-Wesley, 1999.

[2] *MATLAB R2008b*, The Math Works Inc., 2008.

[3] Guoliang Zeng, A New Implementation of Root Raised Cosine Filter with Motorola's DSP56K, *Proceedings of the International Embedded Solutions Event,* September, 2004.

A Computationally Dynamic Resource Management for Multicarrier Communications in Downlink for OFDMA Systems

Kabaou Mohamed Ouwais[1], Rhaimi Belgcem Chibani[1], Abdelkrim Med Naceur[1], Marzoug Mongi[2]

[1]National School of Engineering (ENIG), R.U: Modeling, Analysis and Control Systems (MACS), 6029-Gabès-Tunisia

[2]Orange-Groupe France Telecom, Département Réseaux fixes et Mobiles, 41-45 Bd Romain Rolland, 75672 Paris, Cedex 14, Paris-France

Email: ouwayess@yahoo.fr, kacem_enig4@hotmail.com, abdelkrim.naceur@enig.rnu.tn,

mongi.marzoug@orange-ftgroup.com

Abstract

Multicarrier transmission schemes have been widely adopted in wireless communication systems. Multiple users can share a multicarrier channel using multiple access mechanisms such as OFDMA. This paper studies some unique features of multicarrier transmission and how these effect the performance of a multiple access system. Algorithms are derived for resource allocation in static and dynamic OFDMA systems. Numerical results demonstrate that the proposed approach (algorithm) offer comparable performance with existing algorithms. In this paper, a new algorithm for subcarrier allocation (Hybrid BABS) for the downlink of multiuser OFDM transmission is presented. The proposed algorithm is more stable and it offers a lower complexity and better performance than previous existing algorithms.

Keywords

OFDMA, Resource Management, Bandwidth allocation, Greedy algorithm, Multiuser, Subcarrier, Frequency allocation, Dynamic allocation.

I. INTRODUCTION

In recent years, wireless communications has grown to permeate all facets of life. From cellular access to wireless data networks, there is a demand for clear, fast transmission of multimedia information. In an orthogonal frequency division multiple access (OFDMA) system users are assigned different carriers instead of sharing them. Optimal algorithms for frequency allocation are found to be computationally demanding. In this paper a new class of algorithms is proposed which achieves the users QoS objectives, compared to existing algorithms and at lower computational cost. We study the problem of finding an optimal subcarrier and power allocation strategy for downlink communication to multiple users in an OFDM based wireless system. We formulate the problem of minimizing total power consumption with constraints on BER (Bit Error Rate). Orthogonal Frequency Division Multiple Access (OFDMA) is a promising multiple access technique for next wireless broadband networks. By dividing the bandwidth into multiple subsets of subcarriers (named subchannels), OFDMA exploits the multi-user diversity and thus increases the system throughput.

II. RADIO RESOURCE MANAGEMENT FOR ORTHOGONAL FREQUENCY DIVISION MULTIPLE ACCESS (OFDMA)

One of the biggest advantages of OFDM systems is the ability to allocate power and rate optimally across frequency, using multiple methods (algorithms) for scheduling, management and allocation of radio resource. Computationally efficient algorithms exist to perform management for single or multiple users communications. Kivanc et.al. [1] was one of the researchers to show the problem of finding an optimal sub-carrier and power allocation strategy for downlink communication to multiple users in an OFDM based wireless system. Lengoumbi et.al [2] shows the Rate Adaptive optimization (RA) problem, which maximizes the sum of user data rates subject to total power constraint and individual guaranteed rates. The first algorithm for resource allocation was introduced by Rohling and Grunheid [3]. They present a simple heuristic greedy algorithm, and show that it performs better than simple banded OFDMA. Several problems have been studied,

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_83,

including maximizing the number of users with equal rate requirements [4], maximizing the minimum transmission rate for users with limited power [5].

The main result of this paper is a class of algorithms which have been proposed for subcarrier assignment based on BABS, modified BABS and our Hybrid BABS algorithm. These algorithms achieve comparable performance to the BABS algorithm but do not require intensive computation. A single cell with one base station and many mobile stations is considered. The algorithms assume perfect information about the channel state due to multipath fading as well as path loss and shadowing effects, and the presence of a medium access protocol to convey information about channel state and subcarrier allocation between the base station and the mobile stations.

III. SYSTEM MODEL AND PROBLEM DESCRIPTION

This section introduces the system model that is used, and formally introduces the problem being addressed in this paper. The criteria used for resource allocation are introduced, and the subcarrier allocation method is formulated. We consider the downlink transmissions of a multi-cell OFDMA system.

The system under consideration is an OFDM system with frequency division multiple access (FDMA). Perfect channel state information is assumed at both the receiver and the transmitter, i.e. the channel gain on each subcarrier due to path loss, shadowing, and multipath fading is assumed to be known. Channel parameters are assumed to be perfectly estimated by a channel estimation algorithm for downlink OFDMA [3]. The design and performance of this algorithm is outside the scope of these studies. The system does not employ spreading in either time or frequency, and each subcarrier can only be used by one user at any given time. Subcarrier allocation is performed at the base station and the users are notified of the carriers chosen for them. After the allocation, each user performs power allocation and bit loading across the subcarriers allocated to it to find the transmission power.

We consider a system with K users, and N subcarriers. Each user k must transmit at least R_{k_min} bits per unit time. Let $H_k(n)$ be the channel gain, $p_k(n)$ the transmission power and $r_k(n)$ the transmission rate for user k on subcarrier n.

IV. OPTIMIZATION TECHNIQUES FOR RADIO RESOURCES MANAGEMENT

The objective is to find a subcarrier allocation which allows each user to satisfy its rate requirements. Most famous problems can in fact be described by: [4], [6].

- RA (Rate Adaptive optimization problem),
- MA (Margin Adaptive optimization problem),
- Equity optimization problem,
- Outage optimization problem.

For more details, [2], [7], [9] could be named as several researches having been working about the Rate Adaptive optimization problem (RA). We can also refer to [3], [10], [11] for the Margin Adaptive optimization problem (MA). Doufexi and Armour [14] investigate the problem of dynamic multiuser subcarrier allocation in a coded OFDMA system based on the Equity problem optimization. Finally, for Outage problem optimization we name to [4], [6].

A. Rate Adaptive problem formulation (RA)

Our objective is to optimize bandwidth allocation in the system such that a maximum number of users can achieve their data rate requirements. We aim to find a bandwidth allocation that allows each user to satisfy its rate requirements while maximizing the sum of user data rates subject to total power constraint. Individual guaranteed rates, is considered [8], [10].

The optimization problem can be expressed by:

$$\begin{cases} \max \sum\limits_{u=1..U} r_u & \text{with constraint} \\ \sum\limits_{n=1..N_b} p_{b,n} \le P_{Maxb} & \text{for b=1...B} \end{cases}$$

$$and$$

$$r_u \ge r_u^0$$

Where U is the total number of users, N is the number of OFDMA subcarriers, $P_{tr,max}$ the total power constraint, r and r^0 are respectively the data rate and data rate constraint (minimum data rate) for user u.

V. SENSIBLE GREEDY APPROACH

The problem posed by Wong.et.al [5] is computationally intractable, and as described above, a direct approach to solving it does not yield a good algorithm. This paper examines two different algorithms (resource allocation and subcarrier assignment) and which use information about users channel and rate requirements to find a close approximation to the solution.

Intuitively, the problem is separated into two stages:

1. Resource Allocation: Decide the number of subcarriers each user gets its bandwidth based on rate requirements and the users average channel gain.
2. Subcarrier Allocation: Use the result of the resource allocation stage and channel information to allocate the subcarriers to the users.

A. Resource Allocation Algorithms

In a wireless environment, some users will see a lower overall SNR than other users. These users tend to require the most power. Studying the subcarrier allocations from the LR algorithm shows that once users have enough subcarriers to satisfy their minimum rate requirements, giving more subcarriers to users with lower average SNR helps to reduce the total transmission power. This section describes the bandwidth assignment based on SNR (BABS) algorithm which uses the average SNR for each user to decide the number of subcarriers that user will be assigned. In the second part of this paper, it finds our contribution for the best assignment of subcarriers to users. Two different approaches are presented, the rate-craving greedy algorithm (RCG) and the amplitude-craving greedy algorithm (ACG).

1) Bandwidth Allocation Based on SNR (BABS)

To find the optimal distribution of subcarriers among users given the flat channel assumption, we propose using a greedy descent algorithm similar to that used for discrete water filling.

$$m_k \leftarrow \left\lceil \frac{R_{\min}^k}{R_{\max}} \right\rceil$$

$$while \ \sum_{k=1}^{K} m_k > N \quad do$$

$$k^* \leftarrow \arg \min_{1 \le k \le K} m_k$$

$$m_{k^*} \leftarrow 0$$

$$end \quad while$$

$$while \ \sum_{k=1}^{K} m_k < N \quad do$$

$$G_k \leftarrow \frac{m_k + 1}{H_k} * f\left(\frac{R_{\min}^k}{m_k + 1}\right) - \frac{m_k}{H_k} * f\left(\frac{R_{\min}^k}{m_k}\right), k = 1..K$$

$$l \leftarrow \arg \min_{1 \le k \le K} G_k$$

$$m_l \leftarrow m_l + 1$$

$$end \quad while$$

Where H_k, is a channel gain on every subcarrier, R_{max} is the maximum transmission rate per carrier, and R^k_{min} is a minimum transmission rate requested for user k. Let user k be allocated m_k subcarriers.

Example

- 64 subcarriers,
- 8 users [u_1 u_2... u_8] and [m_1 m_2 ... m_8] is the vector of number of subcarriers allocated for each user.

Before BABS algorithm application, the vector number of subcarriers is: [m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8] = [14 20 11 16 19 21 23 26].

The new vector of number of subcarriers after BABS algorithm application is: [m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8] = [0 0 0 0 0 0 23 26].

B. Subcarrier Assignment Algorithms

Once the number of subcarriers is determined, we move on to assigning specific subcarriers to users. In this section, we propose two suboptimal algorithms to allocate subcarriers to users:

- Amplitude Craving Greedy Algorithm (ACG),
- Rate Craving Greedy Algorithm (RCG).

To minimize the transmission power and allocate the most appropriate subcarriers to the users according to the channel quality information, the amplitude craving greedy (ACG) algorithm is formulated in a mathematical model. The amplitude craving greedy algorithm allocates subcarriers in a random order [2]. The aim of the ACG algorithm is, if the individual users did not have rate constraints to satisfy we would allocate each subcarrier to the user who has the highest gain on that subcarrier. The RCG algorithm begins with an estimate of the users transmission rate on each carrier and aims to maximize the total transmission rate. The ACG algorithm is a modification of RCG which achieves comparable performance at reduced computational complexity. ACG reduces complexity but has more power transmitted than RCG [3].

The following algorithms, represents respectively the ACG and RCG algorithms for subcarrier assignment (Equation.2):

Algorithm 1: ACG

for each subcarrier, $n = 1..N$

$k^* \leftarrow \arg \max_{1 \le k \le K} |H_k(n)|^2$

Let $\#A_k$ denote the cardinality of set A_k

while $\left(\#A_{k^*} = m_{k^*}\right)$ do

$|H_k(n)|^2 \leftarrow 0$

$k \leftarrow \arg \max_{1 \le k \le K} |H_k(n)|^2$

end while

$A_{k^*} \leftarrow A_{k^*} \cup \{n\}$

end for

Where m_k, is the number of subcarriers allocated to each user, and $A_k \leftarrow \{\ \}$ initially for both algorithms.

Algorithm 2: RCG

for each subcarrier, $n = 1..N$

$k^* \leftarrow \arg \max_{1 \le k \le K} r_k(n)$

$A_{k^*} \leftarrow A_{k^*} \cup \{n\}$

end for

for all user, k such that $\#A_k > m_k$, *do*

while $\#A_k > m_k$, *do*

$l^* \leftarrow \arg \min_{\{l:\#A_k < m_l\}} \min_{\{1 \le n \le N\}} -r_k(n) + r_l(n)$

$n^* \leftarrow \arg \min_{\{1 \le n \le N\}} -r_k(n) + r_l(n)$

$A_k \leftarrow A_k / \{n^*\},\ A_{l^*} \leftarrow A_{l^*} \cup \{n^*\}$

end while

end for

where m_k, is the number of subcarriers allocated to each user. $r_k(n)$, is the estimated transmission rate of user k on subcarrier n.

The aim is to find the suboptimum subcarrier allocation algorithm that consumes the minimum transmission power and satisfies each user's rate requirements. This is accomplished in:

$$\begin{cases} \max \sum_{b=1}^{B} \sum_{u=1}^{U} ru,b \\ P_{tr}^{(b)} < P_{tr,\max} (1<u<U), (1<b<B) \end{cases}$$

$$and$$

$$r \ge r^0$$

VI. SIMULATION SETUP

In simulations Bandwidth Assignment Based on SNR (BABS), Modified BABS (M-BABS), and Hybrid BABS (H-BABS) are described. The channel gain $g(u,n)$, of user u on subcarrier n in the serving cell, is summarized as: $g(u;\ n) = K * d(u)^{-\alpha} * a_{sh}(u) * a_f\ (u,n)$ where a_f has a Rayleigh distribution and represents the small scale fading. We consider additive white Gaussian noise (AWGN) which is characterized on each subcarrier by a Gaussian random variable with $\sigma^2 = N_0.W / N$. The system under consideration has parameters given in the following Table. The channel model and traffic source models used in simulations are described in more detail below.

Parameter	Value
W (MHz): Total bandwidth	1
U: Number of users in one cell	8
N: Number of OFDM subcarriers	32
R_{max} (bits): Maximum number of bits on a subcarrier	512
K: Path loss constant	10^{-4}
α: Path loss exponent	2.8
D_{max} (km): Maximum distance between user/BS	5
σ_{sh} (dB): Log-normal standard deviation	8
N_0 : Thermal noise (dBm/Hz)	-174

TABLE I
SYSTEM PARAMETERS

A similar result is developed with the same parameters mentioned in Table I. From our simulations, we obtained the average channel gain versus the time (cf. figure 1).

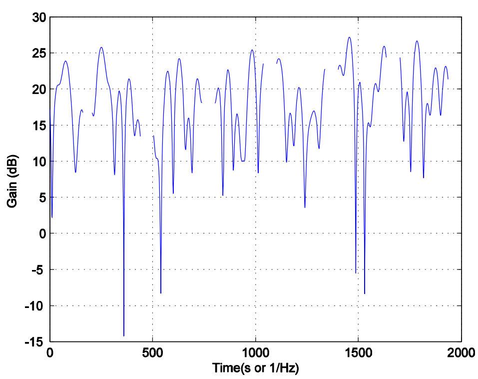


Fig. 1. Channel Gain (dB) versus Time (s)

VII. SIMULATION RESULTS

The algorithms are compared based on three criteria: the efficiency that a BABS algorithm is able to allocate the bandwidth less than Modified BABS (M-BABS) and Hybrid BABS (H-BABS), the optimality, and the data rate per subcarrier offered from each one.

This studies presents performance analysis of resource allocation algorithms for a multicarrier wireless system. The main contributions of this paper are:

- One resource allocation algorithm is proposed: the Hybrid Bandwidth Allocation Based on SNR Algorithm (H-BABS) for decide the number of subcarriers m_k needed for each user. The algorithms are compared using simulations.
- A comparison studies are presented between the two subcarrier allocation algorithms: Amplitude Craving Greedy (ACG) and Rate Craving Greedy (RCG). The main simulation results underlined efficiency and the importance of phase transitions in the study of search suboptimal algorithms.

Figure 2 shows a comparison of the transmitted power, when bandwidth allocation is performed using BABS and modified BABS algorithm. The transmitted power of the BABS algorithm is more important than that of the modified BABS. The requested power is also comparable with the two previous parameters. We conclude that the advantages of the two algorithms are to optimize the transmission power.

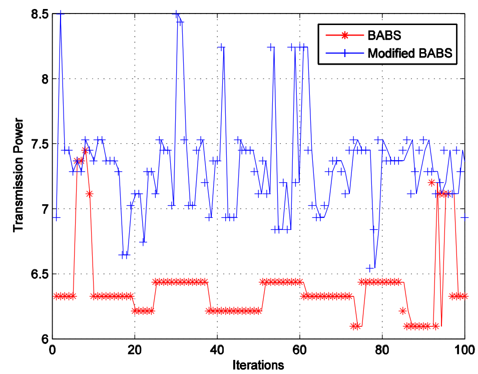


Fig. 2. Transmission power versus Number of Iterations

A. A proposed Hybrid BABS Algorithm for Bandwidth Allocation (H-BABS)

To allocate the bandwidth, we propose an enhancement of BABS. H-BABS is a modified version of BABS and M-BABS, that takes into account the instantaneous algorithms (BABS and H-BABS) during the scheduling. The aim of the proposed algorithm is presented by the following graph:

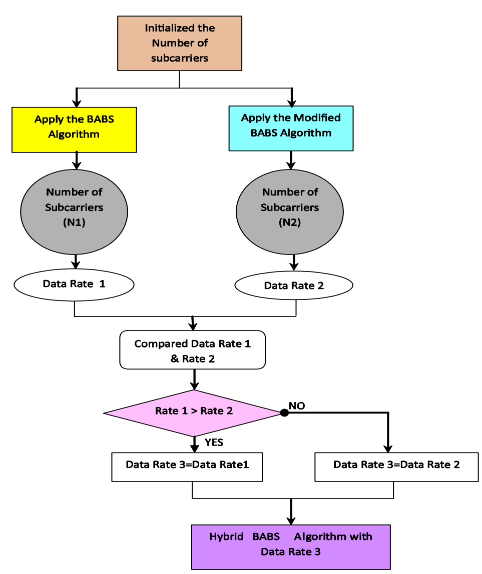


Fig. 3. Hybrid Bandwidth Allocation Based on SNR Algorithm

Figure 4 show the variation in data rate with number of iterations for the BABS, Modified BABS and Hybrid BABS. There are 32 carriers and 8 users in the system. In figure 4, the total bandwidth is W= 1MHz, and every subcarrier is used for data transmission. Figures also show plots of an algorithm labeled as Modified BABS and Hybrid BABS. We observe, that the modified BABS algorithms allow an important gain of data rate than the based BABS algorithm. We can see also in the same figure a clear gap between the two data rates offers by the algorithms. That's why, we can affirmed that the performance obtained by the proposed Hybrid BABS (H-BABS) algorithm is validated.

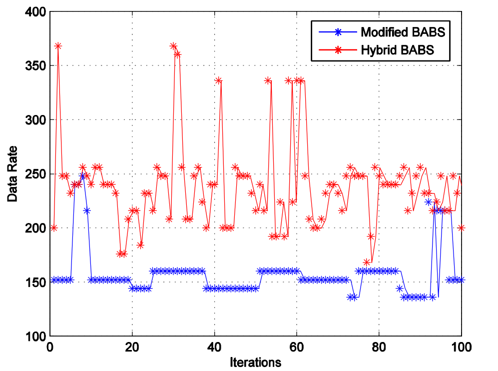


Fig. 4. Comparison between Modified BABS and Hybrid BABS

B. Subcarrier Assignment Algorithms Results

Figure 5 gives a closer look at the BABS-ACG, BABS-RCG, algorithms for the typical channel with the gain mentioned by figure1. The plot for BABS-ACG and BABS-RCG are nearly adjacent and is difficult to see the difference between both. The BABS-ACG algorithm searches all possible allocations, and yields the lowest transmission power per bit.

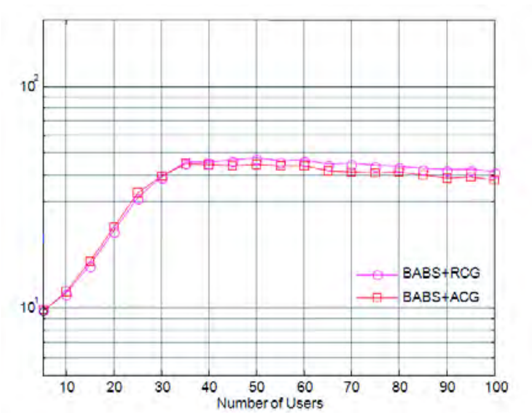


Fig. 5. Power per bit Transmitted ($P_T = R_{tot}$)

VIII. Conclusion

In this paper, we have designed a radio resource management scheme that integrates dynamic subcarrier assignment in OFDMA wireless networks. We have formulated the optimal subcarrier assignment scheme, and proposed a new suboptimal algorithm for resource allocation. Our simulations have demonstrated that RRM can more efficiently utilize the available subcarriers. In this paper, a computationally efficient class of algorithms for allocating subcarriers among users in a multicarrier system has been described. Dividing the problem into two stages: resource allocation and subcarrier assignment, we present the main objective of each sub-problem, the simulation results underlined the opportunity of the approach. This approach allows efficient use of system resources in terms of bandwidth efficiency.

REFERENCES

[1] Didem. Kivanc and Hui. Liu, *Subcarrier Allocation and Power Control for OFDMG*, IEEE Transactions on Wireless Communications, November 2003, vol. 2, no. 6, pp. 1150-1158.

[2] Carle. Lengoumbi, Philippe. Godlewski, Philippe. Martins, *An Efficient Subcarrier Assignment Algorithm for Downlink OFDMA,* The 64th IEEE Semiannual Vehicular Technology Conference, Montreal, Canada, 25-28 september 2006, pp. 1010-1020.

[3] Didem Kivanc. Tureli, *Resource Allocation for Multicarrier Communications*, Doctorat of Philosophy, University of Washington, 2005.

[4] Carle. Lengoumbi, *Accès multiple OFDMA pour les systémes cellulaires post 3G: Allocation de ressources et ordonnancement*, Thèse Doctorat, Ecole Nationale Supérieur des Télécommunications (ENST Paris), Décembre 2008.

[5] C.Y.Wong, R.S. Cheng, K.B. Letaief, and R.D. Much., *Multiuser OFDM with adaptive subcarrier, bit and power allocation*, IEEE Journal. on Sel, Areas in Communmmunication, pp. 17(10):1747-1757, October 1999.

[6] Hammouda. Soumaya, *Analyse et amélioration de la gestion des ressources radio du système radio mobile de troisième génération* UMTS *et au-delà*, Thèse doctorat, Ecole Supérieure des communications de Tunis (SUP'COM), Technologie de l'information et de la communication, Juin 2007.

[7] Carle. Lengoumbi, Philippe. Godlewski, Philippe. Martins, *Dynamic subcarrier reuse with rate guaranty in a downlink multicell OFDMA system*, The 17th Annual IEEE International

Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'06), 2006.

[8] Jiho. Jang, Kwang Bok. Lee, *Transmit Power Adaptation for Multiuser OFDMA Systems*, IEEE Journal on selected areas in communications, 21. NO 2, pp. 171- 178, February 2003.

[9] Tommy K. Chee, Cheng-Chew. Lim, Jinho. Choi, *Suboptimal Power Allocation for Downlink OFDMA Systems*, IEEE 60th VTC 2004, 26-29 September 2004.

[10] D. Kivanc, H. Liu, *Subcarrier Allocation and Power Control for OFDMA*, Conference Record of the Thirty-Fourth Asilomar Conference on Signals, Systems, and Computers, pp. 147 -151, 2006.

[11] Inhyoung. Kim, Hae Leem. Lee, Beomsup. Kim, and Yong H. Lee, *On the Use of Linear Programming for Dynamic Sub channel and Bit Allocation in Multiuser OFDM*, In Proc. IEEE Global Telecommunications Conf., volume 6, pp. 3648 -3652,2001.

[12] Lei. Yan, *An introduction to Bandwidth Allocation in OFDM*, March 2008.

[13] Carle. Lengoumbi, Philippe. Godlewski, Philippe. Martins, *Comparison of different subchanalization modes for OFDMA*, The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07), 2007.

[14] A. Doufexi, S. Armour., *Design considerations and Physical Layer Performance Results for a 4G OFDMA System Employing Dynamic Subcarrier Allocation*, The 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2005.

A General Method for Synthesis of Families of Orthogonal Complementary Codes

B. Y. Bedzhev, Member, IEEE
University of Shumen “Bishop Konstantin Preslavsky”
115 Universitetska Str.
Shumen, Bulgaria 9712, EU
M. P. Iliev
University of Ruse “Angel Kanchev”
8 Studentska Str.
Ruse, Bulgaria 7017, EU

***Abstract*-Families of orthogonal complementary codes (FOCCs) are called families of sets of radio signals possessing both zero autocorrelation functions (ACFs) (except for the zero time shift) and zero cross-correlation functions (CCFs) among all pairs of the members of a family. They are viewed as promising tool for a radical enhancement of the data transmission rate of the future wireless communication systems, because they allow the users to communicate practically without multiple access interferences in a multipath environment. Due to this reason in the paper a general method for synthesis of FOCCs is proposed. It allows the parameters of the FOCCs to be independently managed and the most part of techniques for synthesis of FOCCs, known today, to be viewed in a common theoretical frame.**

***Index Terms* – Synthesis of signals, Families of orthogonal complementary codes.**

I. INTRODUCTION

The wireless communication systems are in very rapid progress today. Despite of this the users and industry expect an even more significant enhancement of the quality of the services and the transmission rate, offered by these systems. In order to meet these expectations, the experience, obtained during the past twenty years, has been analyzed by the theorist and developers. The main conclusion is that the critical role for the whole performance of the wireless communication systems has the system of employed radio signals. Due to this reason great efforts have been directed to finding families of radio signals possessing both autocorrelation functions (ACFs) with a small level of the side-lobes and small cross-correlation functions (CCFs) among all pairs of members of a family. The focusing of the research activities on this problem can be explained by the following facts [1], [3], [4], [5], [8], [9], [10], [11], [12], [13], [14], [15]. First, the pointed out families of radio signals allow the so-named self-interference (SI), caused by multipath spreading of electromagnetic waves, to be reduced by a separate processing of the direct and reflected signals. Second, it is possible the negative effect of simultaneous transmission of numerous users, named multi user interference (MAI), to be minimized.

With regard to above described situation, our paper aims to develop a general method for synthesis of families of orthogonal complementary codes (FOCCs), which will allow the most part of techniques for synthesis of FOCCs, known today, to be viewed in a common theoretical frame. The importance of the studied problem results from the following facts. First, today the FOCCs are the only known radio signals that have perfect (both periodic and aperiodic) correlation properties [1], [3], [4], [5], [8], [9], [10], [11], [12], [13], [14], [15]. This means that every member of a FOCC has an ideal ACF, resembling the Dirac delta – function, and the CCFs among all pairs of members of a FOCC are zero for every time shift. Second, despite of taken efforts many problems in the field of the synthesis of the FOCCs are open still.

Paper is organized as follows. First, the basics of the FOCCs are recalled. Second, our method for synthesis of FOCCs is suggested. At the end some important conclusions are given.

II. BASICS OF THE SYNTHESIS OF THE FAMILIES OF ORTHOGONAL COMPLEMENTARY CODES

Let $K \geq 2, M \geq 2, N \geq 1,$ be integers and let us consider a family $V(K, M, N)$, comprising K sets as every set consists of M sequences with (code) length N:

$$\begin{aligned}&\{a_k(m,n)\}_{n=0}^{N-1} = \\ &= \{a_k(m,0), a_k(m,1), ..., a_k(m,N-1)\}, m = 0,1,...,M-1\end{aligned} \quad (1)$$

Then $V(K, M, N)$ is called family of orthogonal complementary codes (FOCCs) if the following conditions are satisfied [4], [7], [9]. First, the aggregated ACFs of the sequences of every set are zero except for the zero time shift. Second, the aggregated CCFs of the sequences of all pairs of sets are zero for every time shift.

These conditions can be mathematically described as follow:

$$R_{A_k,A_l}(r) = \begin{cases} M.N, & k = l \cap r = 0; \\ 0, & k \neq l \cup r \neq 0, \end{cases} \quad (2)$$

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_84, © Springer Science+Business Media B.V. 2010

where:

- r is the time shift, $r = -(N-1),(N-2),...,-1,0,1,...,N-1$;
- A_k, $k = 0,1,...,K-1$ is the matrix, denoting the k-th set of the family $V(K, M, N)$, i.e.:

$$A_k = \begin{bmatrix} a_k(0,0) & a_k(0,1) & ... & a_k(0,N-1) \\ a_k(1,0) & a_k(1,1) & ... & a_k(1,N-1) \\ ... & ... & ... & ... \\ a_k(M-1,0) & a_k(M-1,1) & ... & a_k(M-1,N-1) \end{bmatrix} \quad (3)$$

- $R_{A_k,A_l}(r)$ is the aggregated CCF (ACF if $k = l$) of the rows of the k-th and l-th sets (matrices), $k, l = 0,1,...,K-1$.

The aggregated $R_{A_k,A_l}(r)$ CCF (ACF if $k = l$) is evaluated by the well known formula [1], [2], [3], [4], [13], [14], [15]:

$$R_{A_k,A_l}(r) = \begin{cases} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1-|r|} a_k^*(m,n).a_l(m,n+|r|), & -(N-1) \le r \le 0, \\ \sum_{m=0}^{M-1} \sum_{n=0}^{N-1-r} a_k(m,n).a_l^*(m,n+r), & 0 \le r \le N-1. \end{cases} \quad (4)$$

Here the symbol "*" means "complex conjugation".

It should be mentioned that FOCCs are mathematical models of the radio signals for the multicarrier code-division multiple access (MC-CDMA) wireless communication systems. In such a system K user can simultaneously communicate practically without SI and MAI. The k-th set (i.e. the matrix A_k, $k = 0,1,...,K-1$) of the exploited FOCC is assigned to the k-th user as its signature code. The entry $a_k(m,n)$, $m = 0,1,...,M-1$, $n = 0,1,...,N-1$ of the matrix A_k, $k = 0,1,...,K-1$ is the complex envelope of the n-th elementary pulse of the m-th sequence. In fact the m-th sequence describes mathematically the law of the amplitude and phase modulation of the m-th subcarrier frequency during the every clock interval in the process of the spreading of the spectrum of the radio signals. Most often only phase manipulation (Q-phase shift keying) is used, i.e.

$$\forall a_k(m,n) \in \{\exp[(2\pi i l)/Q; l = 0,1,...,Q-1\}, \; i = \sqrt{-1}, \quad (5)$$

but this restriction will not be used in the sequel.

For brevity but without loss of generality we shall suppose that $0 \le r \le N-1$ in (4). Then, after changing the order of summations, (4) can be transformed as follows:

$$R_{A_k,A_l}(r) = \sum_{n=0}^{N-1-r} \sum_{m=0}^{M-1} a_k(m,n).a_l^*(m,n+r) = \sum_{n=0}^{N-1-r} CA_k(n) \otimes CA_l^*(n+r). \quad (6)$$

Here $CA_k(n)$ is the n-th vector – column of the matrix A_k and the symbol "$\otimes$" denotes the inner product of the n-th vector – column of the matrix A_k and the complex conjugated (n+r)-th vector – column of the matrix A_l. Now it should be recalled that the evaluation of the CCFs can be simplified by means of multiplication of polynomials [2]. More specifically, let us introduce the following polynomials

$$\begin{aligned} F_k(x) &= \sum_{n=0}^{N-1} CA_k(n).x^n, \; k = 0,1,...,K-1, \\ F_l^*(x^{-1}) &= \sum_{j=0}^{N-1} CA_l^*(j).x^{-j}, \; l = 0,1,...,K-1 \end{aligned}, \quad (7)$$

then the coefficients of their polynomial product $F_k(x).F_l^*(x^{-1})$ will be the aggregated cross-correlations $R_{A_k,A_l}(r)$ for all time shifts $r = -(N-1), -(N-2),...,-1,0,1,...,N-1$ according to (6).

With regard to (7), the conditions (2) can be rewritten as follows

$$F_k(x).F_l^*(x^{-1}) = \begin{cases} M.N, & k = l; \\ 0, & k \ne l. \end{cases} \quad (8)$$

It should be stressed that multiplication of the coefficients in (8) is performed as multiplication of vectors, i.e.

$$CA_k(n) \otimes CA_l^*(n+r) = \sum_{m=0}^{M-1} a_k(m,n).a_l^*(m,n+r) \quad (9)$$

and due to this reason it is possible $CA_k(n) \otimes CA_l^*(n+r) = 0$.

III. A General Method for Synthesis of Families of Orthogonal Complementary Codes

The complementary codes were invented by M. Golay in 1949. In his origin works [6], [7] K = 1, M = 2 and $\forall a(m,n) \in \{-1,+1\}$, $m = 0,1$, $n = 0,1,...,N-1$, which gave the reason these codes to be named complementary pairs. Later, in 1972, Tseng and Liu [13] extended the Golay's conception and introduced the so - named mutually orthogonal complementary sets, consisting of $M \ge 2$ complementary sequences.

Due to their unique correlation properties the complementary codes have been intensively studied. For instance the study in [4] is based on over 300 journal papers and conference reports, concerning this theme. The complementary codes have also exploited in some types radars and specialized communication systems and obtained experience has proved the practical effectiveness of these signals. As a result at the beginning of the new century it was suggested the FOCCs, described in the previous section, to be the base tool for multiple access to the communication channel in the next generations wireless communication systems [1], [3], [4], [5], [8], [9], [12], [14], [15].

At the moment despite of the taken efforts in the study of complementary codes, it is not known if or not the parameters *K*, *M*, *N* of a FOCC can be independently managed during the developing of a new wireless communication system. For example in [3], [4], [9], [13], [14] methods for synthesis of FOCCs are described, according to which an augmentation of the number of users (i.e. *K*) is possible only by enlarging of the length *N* of the sequences. Unfortunately, this approach leads to decreasing of the data transmission rate, which is very undesirable. Due to this reason in this part of the paper a general method for synthesis of FOCCs is suggested. In allows the parameters *K*, *M*, *N* of a FOCC to be chosen independently during the process of system optimization. Besides our method provides the opportunity a significant part of known today techniques for synthesis of FOCCs to be viewed from a common theoretical base.

The suggested general method for synthesis of GOCCs is based on the following propositions.

Proposition 3.1: Let *A* be a set of *M* complementary sequences with length *N* and let *D* be a matrix

$$D=\begin{bmatrix} d(0,0) & d(0,1) & \dots & d(0,K-1) \\ d(1,0) & d(1,1) & \dots & d(1,K-1) \\ \dots & \dots & \dots & \dots \\ d(K-1,0) & d(K-1,1) & \dots & d(K-1,K-1) \end{bmatrix} \quad (10)$$

which columns are mutually orthogonal, i.e.

$$\sum_{k=0}^{K-1} d(k,n).d^*(k,l)=\begin{cases} K, & n=l; \\ 0, & n\neq l, \end{cases} \quad (11)$$

for $0\le n,l\le K-1$. Then the column - submatrices of the tensor product

$$E=\begin{bmatrix} d(0,0).A & d(0,1).A & \dots & d(0,K-1).A \\ d(1,0).A & d(1,1).A & \dots & d(1,K-1).A \\ \dots & \dots & \dots & \dots \\ d(K-1,0).A & d(K-1,1).A & \dots & d(K-1,K-1).A \end{bmatrix} \quad (12)$$

are a FOCC with parameters: family size – *K*, number of sequences in every set – *M'*= *K.M* and length of the sequences – *N*.

Proof: According to (7) let $F(x)$ and $F^*(x^{-1})$ be the polynomials of (*N*-1)-th degree which coefficients are the vector – columns of the matrix *A*. As *A* is a set of *M* complementary sequences with length *N*:

$$F(x).F^*(x^{-1})=M.N \; . \quad (13)$$

Let $CE(n)$ be the *n*-th column – submatrix of the matrix *A*:

$$CE(n)=\left[d(0,n).A \quad d(1,n).A \quad \dots \quad d(K-1,n).A\right]^T , \quad (14)$$

where the symbol “T” denotes “transposition”.

After taking into account (11) and (13), the CCF of the *n*-th and *l*-th column - submatrices $CE(n)$ and $CE(j)$ of the tensor product (12) can be evaluated as follows

$$\begin{bmatrix} d(0,n).F(x) \\ d(1,n).F(x) \\ \dots \\ d(K-1,n).F(x) \end{bmatrix} \otimes \begin{bmatrix} d^*(0,l).F^*(x^{-1}) \\ d^*(1,l).F^*(x^{-1}) \\ \dots \\ d^*(K-1,l).F^*(x^{-1}) \end{bmatrix} =$$

$$=\sum_{k=0}^{K-1}\left[d(k,n).F(x)\right].\left[d^*(k,l).F^*(x^{-1})\right]= \quad (15)$$

$$=M.N\sum_{k=0}^{K-1} d(k,n).d^*(k,l)=\begin{cases} (K.M).N, & n=l \\ 0, & n\neq l. \end{cases}$$

Equation (15) proves the Proposition 3.1.

It should be emphasized on the following facts.

First, the rows of *D* are mutually orthogonal also, because *D* is a square matrix [1].

Second, the Proposition 3.1 remains true if we reduce some of columns of *D* and use only *K'* (*K'*<*K*) columns for creating the tensor product (12). Anyway in this case the size of the synthesized FOCC will be smaller – namely, it will be *K'*.

Third, on the base of a known initial set of *M* complementary sequences with length *N* the Proposition 3.1 proves a technique for synthesis of a FOCC with parameters: family size – *K*, number of sequences in every set – *M'*= *K.M* and length of the sequences – *N*.

Proposition 3.2: Let *V*(*K*, *M*, *N*) be a FOCC consisting of *K* sets $A_0, A_1, \dots, A_{K-1}$ as every set comprises *M* complementary sequences with length *N*. Besides, let *D* be a matrix, which columns are orthogonal, according to (10) and (11). Then the column - submatrices of the all tensor products

$$E_k = \begin{bmatrix} d(0,0).A_{u(k,0)} & \dots & \dots & d(0,K-1).A_{u(k,0)} \\ d(1,0).A_{u(k,1)} & \dots & \dots & d(1,K-1).A_{u(k,1)} \\ \dots & \dots & \dots & \dots \\ d(K-1,0).A_{u(k,K-1)} & \dots & \dots & d(K-1,K-1).A_{u(k,K-1)} \end{bmatrix}, \quad (16)$$

$$k = 0,1,\dots,K-1$$

are a FOCC with parameters: family size – $K' = K^2$, number of sequences in every set – $M' = K.M$ and length of the sequences – N.

Here the all rows of the matrix U

$$U = \begin{bmatrix} u(0,0) & u(0,1) & \dots & u(0,K-1) \\ u(1,0) & u(1,1) & \dots & u(1,K-1) \\ \dots & \dots & \dots & \dots \\ u(K-1,0) & u(K-1,1) & \dots & u(K-1,K-1) \end{bmatrix} \quad (17)$$

are permutations of the set $\{0,1,\dots,K-1\}$, i.e. $\{u(k,0),u(k,1),\dots,u(k,K-1)\}$ are the numbers $\{0,1,\dots,K-1\}$ in a possibly changed order.

Proof: According to (7) let $F_{k,m}(x)$ and $F^*_{l,m}(x^{-1})$ be the polynomials of (N-1)-th degree which coefficients are the vector – columns of the matrices $A_{u(k,m)}$ and $A^*_{u(l,m)}$ respectively. As all matrices $A_{u(k,m)}$ are sets of M complementary sequences with length N:

$$F_{k,m}(x).F^*_{l,m}(x^{-1}) = \begin{cases} M.N, & k = l, \\ 0, & k \neq l, \end{cases} \quad (18)$$

for $0 \le k,l \le K-1$.

Let $CE_k(n)$ be the n-th column – submatrix of the matrix E_k, i.e.:

$$CE_k(n) = \\ = \begin{bmatrix} d(0,n).A_{u(k,0)} & \dots & \dots & d(K-1,n).A_{u(k,K-1)} \end{bmatrix}^T, \quad (19)$$

where the symbol “^{T}” denotes “transposition”.

After taking into account (11), (18) and (19), the CCF of the n-th column – submatrix $CE_k(n)$ of the k-th tensor product and j-th column - submatrix $CE_l(j)$ of the l-th tensor product (16) can be evaluated as follows

$$\begin{bmatrix} d(0,n).F_{k,0}(x) \\ d(1,n).F_{k,1}(x) \\ \dots \\ d(K-1,n).F_{k,K-1}(x) \end{bmatrix} \otimes \begin{bmatrix} d^*(0,j).F^*_{l,0}(x^{-1}) \\ d^*(1,j).F^*_{l,1}(x^{-1}) \\ \dots \\ d^*(K-1,j).F^*_{l,K-1}(x^{-1}) \end{bmatrix} =$$

$$= \sum_{m=0}^{K-1} \left[d(k,n).F_{k,m}(x)\right]\left[d^*(k,j).F^*_{l,m}(x^{-1})\right] = \quad (20)$$

$$= \begin{cases} (K.M).N, & k = l \cap n = j \\ 0, & k \neq l \cup n \neq j. \end{cases}$$

Equation (20) proves the Proposition 3.2.

It should be pointed out that the Proposition 3.2 suggests a technique for expanding K times the size of an initial FOCC without changing the length of the sequences. This is a very useful opportunity, because the known today methods [3], [4], [9], [13], [14] expand the size of an initial FOCC only by a simultaneous enlargement of the length of the sequences. Anyway, in order to show that parameters of a FOCC can be independently managed it is necessary to prove the following propositions.

Proposition 3.3: Let $V(K, M, N)$ be a FOCC, consisting of K sets $A_0, A_1, \dots, A_{K-1}$ as every set comprises M complementary sequences with length N. Besides, let D be a matrix which columns are orthogonal, according to (10) and (11). Then the row - submatrices of the tensor product

$$H = \begin{bmatrix} d(0,0).A_0 & d(0,1).A_1 & \dots & d(0,K-1).A_{K-1} \\ d(1,0).A_0 & d(1,1).A_1 & \dots & d(1,K-1).A_{K-1} \\ \dots & \dots & \dots & \dots \\ d(K-1,0).A_0 & d(K-1,1).A_1 & \dots & d(K-1,K-1).A_{K-1} \end{bmatrix} \quad (21)$$

are a FOCC with parameters: family size – K, number of sequences in a set –M and length of the sequences – $N' = K.N$.

Proof: According to (7) let $F_k(x)$ and $F^*_k(x^{-1})$ be the polynomials of (N-1)-th degree which coefficients are the vector – columns of the matrix A_k. As $A_0, A_1, \dots, A_{K-1}$ are a FOCC with parameters K, M, N:

$$F_k(x).F^*_l(x^{-1}) = \begin{cases} M.N, & k = l, \\ 0, & k \neq l, \end{cases} \quad (22)$$

for $0 \le k,l \le K-1$.

Let $RH_k(m)$ be the m-th row – submatrix of the tensor product H:

$$CH_k(m) = \left[d(m,0).A_0 \quad d(m,1).A_1 \quad \dots \quad d(m,K-1).A_{K-1}\right]. \quad (23)$$

After taking into account (22) and (23), the CCF of the m-th and l-th row - submatrices $CH_k(m)$ and $CH_l(m)$ of the tensor product (21) can be evaluated as follows

$$\begin{aligned}
&\left[\sum_{k=0}^{K-1} d(m,k).F_k(x).x^{k.N}\right]\left[\sum_{j=0}^{K-1} d^*(l,j).F_j^*(x^{-1})x^{-j.N}\right] = \\
&= \sum_{k=0}^{K-1} d(m,k).d^*(l,k).F_k(x).F_k^*(x^{-1}) + \\
&+ \sum_{k=0}^{K-1}\sum_{\substack{j=0,\\ j\neq k}}^{K-1} d(m,k).d^*(l,j).F_k(x).F_j^*(x^{-1}).x^{(k-j).N} = \\
&= M.N\sum_{k=0}^{K-1} d(m,k).d^*(l,k) + 0 = \begin{cases} M(K.N), & m = l \\ 0, & n \neq l. \end{cases}
\end{aligned} \quad (24)$$

Equation (24) proves the Proposition 3.3.

Proposition 3.4: After interleaving of the columns of the row – submatrices $CH_k(m)$, defined by (23), a FOCC with parameters: family size – K, number of sequences in a set –M and length of the sequences – $N' = K.N$ will be obtained.

The Proposition 3.4 can be proven in a truly similar way as above and due to this reason the proof is left to the reader.

IV. Results

It should be pointed out that the Proposition 3.3 and 3.4 suggest techniques for expanding K times the length of the sequences of an initial FOCC without changing the number of the sequences. Consequently, Propositions 3.2, 3.3 and 3.4 together prove that the parameters of the FOCCs can be independently managed.

The Propositions 3.1, 3.2 and 3.3 will be explained by the following example.

Example 4.1: Let us make use of the simplest set of complementary sequences

$$A = [+1] \quad (25)$$

with parameters $K = M = N = 1$ and the simplest orthogonal matrix

$$D = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}. \quad (26)$$

After applying the Proposition 3.1 we obtain the FOCC

$$A'_0 = \begin{bmatrix} +1 \\ +1 \end{bmatrix}, \; A'_1 = \begin{bmatrix} +1 \\ -1 \end{bmatrix} \quad (27)$$

with parameters $K' = 2$, $M' = K.M = 2$ and $N' = 1$.

Now, using the Proposition 3.3, FOCC from (27) and D, defined by (26), we create the following FOCC

$$A''_0 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}, \; A''_1 = \begin{bmatrix} +1 & -1 \\ +1 & +1 \end{bmatrix} \quad (28)$$

with parameters $K'' = 2$, $M'' = 2$ and $N'' = K.N' = 2$.

At the end, employing the simplest matrix U, which rows are the numbers $\{0,1,\dots,K-1\}$ in their natural order (here $K = 2$)

$$U = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \quad (29)$$

and applying the Proposition 3.2, FOCC from (28), again D, defined by (26), we create the following FOCC

$$\begin{aligned}
A'''_0 &= \begin{bmatrix} +1 & +1 \\ +1 & -1 \\ +1 & +1 \\ +1 & -1 \end{bmatrix}, \; A'''_1 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \\ -1 & -1 \\ -1 & +1 \end{bmatrix}, \\
A'''_2 &= \begin{bmatrix} +1 & -1 \\ +1 & +1 \\ +1 & -1 \\ +1 & +1 \end{bmatrix}, \; A'''_3 = \begin{bmatrix} +1 & -1 \\ +1 & +1 \\ -1 & +1 \\ -1 & -1 \end{bmatrix}
\end{aligned} \quad (30)$$

The parameters of the FOCC, presented by its sets of complementary sequences (30), are: $K''' = K.K'' = 4$, $M''' = K.M'' = 4$ and $N''' = N'' = 2$.

It should be stressed that the analogous FOCC, synthesized by the methods, known from [3], [4], [13], [14] is

$$\begin{aligned}
B_0 &= \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & +1 & -1 & -1 \\ -1 & +1 & -1 & +1 \\ -1 & +1 & +1 & -1 \end{bmatrix}, \; B_1 = \begin{bmatrix} -1 & -1 & +1 & +1 \\ -1 & -1 & -1 & -1 \\ +1 & -1 & -1 & +1 \\ +1 & -1 & +1 & +1 \end{bmatrix}, \\
B_2 &= \begin{bmatrix} -1 & +1 & -1 & +1 \\ -1 & +1 & +1 & -1 \\ +1 & +1 & +1 & +1 \\ +1 & +1 & -1 & -1 \end{bmatrix}, \; B_3 = \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & -1 & +1 & -1 \\ -1 & -1 & +1 & +1 \\ -1 & -1 & -1 & -1 \end{bmatrix}
\end{aligned} \quad (31)$$

As seen from (30) and (31), the classical methods for synthesis of FOCCs extend the size of a FOCC only by the cost of a directly proportional enlargement of the length of the sequences, used for spreading the spectrum of the radio signals. This peculiarity often has negative effect over the performance of the developed wireless communication system as it leads to an undesirable decrease of the data transmission rate. In contrast, the general method for synthesis of FOCCs, suggested in this part of the paper, allows the size of the family and the length of the sequences of a FOCC to be independently managed.

V. Conclusion

In the paper a general method for synthesis of FOCCs is suggested. It is based on the Propositions 3.1, 3.2, 3.3 and 3.4, proved in paper.

It should be emphasized on the following facts.

First, the Proposition 3.1 provides a technique for synthesis of an initial FOCC with parameters: family size – K, number of sequences in a set – $M' = K.M$ and length of the sequences – N, if an initial set of M complementary sequences with length N is known. Methods for synthesis of initial sets of complementary sequences are presented in [1], [4], [7], [8], [13].

Second, the Propositions 3.3 and 3.4 allow the length of the sequences of a FOCC to be enlarged if it is necessary in order to expand the spectrum of the employed signals. Here it ought to be recalled that the spreading of the spectrum of the signals provides high processing gain (i.e. increasing the signal-to-noise ratio) in receivers of the developed wireless communication system.

Third, the Proposition 3.2 substantiates a technique for expanding K times the size of a FOCC without changing the length of the sequences. Here it ought to be stressed that family size of a FOCC is very important parameter as it determines the number of users, who can simultaneously exploit the resources of the wireless communication system without MAI. In contrast the classical methods for synthesis of FOCCs extend the size of the FOCC only by the cost of a directly proportional enlargement of the length of the sequences. This "hard" connection between the family size and the length of the FOCC could negatively influence the process of development of the new wireless communication systems, because often leads to an undesirable diminishing of the data transmission rate.

Forth, the Example 3.5 shows that if we flexibly apply the techniques, provided by Propositions 3.1, 3.2, 3.3 and 3.4, we can obtain the most part of the FOCCs, known today. For instance, it is obvious that the columns of the orthogonal matrix D, defined by (10) and (11), is a FOCC with parameters: $K = M$ and $N = 1$. Then using this FOCC, Propositions 3.3 and the same matrix D, we can easily obtain the constructions, presented in [1], [4], [8], [13], [14]. Moreover, if we use the Proposition 3.4 instead of the Proposition 3.3, we shall obtain a new construction, which is not covered by [1], [4], [5], [8], [12], [13], [14], [15].

With regard to all above stated we hope the general method for synthesis of FOCCs, suggested in the paper, to be useful in the process of development of the new generation wireless communication system, which aim to provide a radical enhancement of the data transmission rate.

Acknowledgment

The authors wish to thank Prof. Dr Kunchev, Prof. Dr E. Ferdinandov, Prof. Dr H. Hristov and Prof. Dr B. Tsankov. Finally, the authors appreciate the thoughtful critique by the referees.

References

[1] B. Y. Bedzhev, "Generalized orthogonal complementary codes for CDMA wireless communications", in Proc. *The 2004 European Conference on Wireless Technology*, pp. 329–332, Amsterdam, Kingdom of Netherlands, Oct. 2004.

[2] R .E. Blahut, *Theory and Practice of Error Control Coding*, Addison-Wesley, Reading, Mass., USA, 1987

[3] H.-H. Chen, J.-F. Yeh and N. Suehiro, "A multicarrier CDMA architecture based on orthogonal complementary codes for new generations of wideband wireless comunicatios," *IEEE Communications Magazine*, pp. 126–135, Oct. 2001.

[4] H.-H. Chen, *Next Generation CDMA Technologies*, John Wiley&Sons, Chichester, West Sussex, England, 2007

[5] A. T. Fam and I. Sarkar, "A new class of interlaced complementary codes based on components with unity peak sidelobes," *Signal Processing*, vol. 88 (2008), pp. 307–314

[6] M. J. E. Golay, "Multislit spectroscopy," *J. Opt. Soc. Amer.*, vol. 39, pp. 437–444, 1949

[7] M. J. E. Golay, "Complementary series," *IRE Trans. Inform. Theory*, vol. 7, pp. 82–87, №2, 1961.

[8] V. V. Ignatov, S. A. Dobrovolsky and A. Y. Guzhva, "Matrix systems of signals for the CDMA networks", *Elektrosvyaz*, pp. 41-42, Sept. 2003 (in Russian).

[9] Y. Jin and H. Koga, "Basic properties of the complete complementary codes using the DFT matrices and the Kronecker products", in Proc. *International Symposium on Information Theory and its Applications, ISITA2008*, pp. 1-6, Auckland, New Zealand, 7-10 December, 2008

[10] S. P. Maity, M. Mukherjee, "A high capacity CI/MC-CDMA system for reduction in PAPR," in Proc. *Seventh Annual Communication Networks and Services Research Conference*, pp.297-304, Moncton, New Brunswick, Canada, 11-13 May, 2009

[11] A. Milev, V. Mutkov and B. Najdenov, "An algorithm for construction of optical orthogonal codes," unpublished

[12] T. R. Qureshi, M. D. Zoltowski and R. Calderbank, "A MIMO-OFDM channel estimation scheme utilizing complementary sequences," in Proc. *The 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.2677-2680, Taipei, Taiwan, 19-24 April, 2009

[13] C.-C. Tseng and C. L. Liu, "Complementary set of sequences," *IEEE Trans. Inf. Theory*, vol. 18, pp. 206–209, Sept. 1972.

[14] S.-M. Tseng and M. R. Bell, "Asynchronous multicarrier DS-CDMA using mutually orthogonal complementary sets of sequences," *IEEE Trans. on Communications*, vol. 48, pp. 53–59, Jan. 2000.

[15] M. D. Zoltowski, T. R. Qureshi and R. Calderbank, "Complementary codes based channel estimation for MIMO-OFDM systems," in Proc. *Forty-Sixth Annual Allerton Conference*, pp. 133-138, Allerton House, UIUC, Illinois, USA, 23-26 Sept., 2008

E-Speed Start- A Window Based Additive Increase, Multiplicative Decrease Network Congestion Control Technique

Oyeyinka I. K[1], Akinwale A.T[2], Oluwatope A O[3]. Folorunso O[2], Aderounmu G. A[3], Abiona O. O[4]

[1] Computer Technology Department, Yaba College of Technology, Yaba, Lagos-Nigeria.
ikoyeyinka@yabatech.edu.ng

[2] Computer Science Department, University of Agriculture, Abeokuta, Nigeria.
{atakinwale, folorunsolusegun}@yahoo.com

[3]Comnet Lab., Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife. Nigeria
{aoluwato, gaderoun}@oauife.edu.ng

[4]Department of Computer Information System. Indiana University, NorthWest, USA
oabiona@iun.edu

Abstract: The two major algorithms in the Additive Increase, Multiplicative Decrease Network Congestion Control Algorithm (AIMD) are the slow start and the congestion avoidance. Much works has been done on improving the congestion avoidance stage. Not until recently has attention shifted to slow start. The proposals for improving slow start include swift start, quick start etc. These modifications are deficient in one way or the other. Hence we propose e-speed start for AIMD. We attempt to incorporate environmental factors into slow start speed determination. Furthermore, we propose multiple selectable startups for AIMD. Hence a connection may use the conventional slow start, e-speed start or any other startup algorithm depending on the size of the available bandwidth. In fact, we are of the view that AIMD slow start may be selectable from n-ary set of algorithms. The e-speed start uses the available bandwidth to calculate a factor β which is used to determine the congestion window (cwnd) suitable for transmission at that point. The result obtained shows that e-speed start performs far better than the slow start where the available bandwidth is high and the environmental factors are favourable. It was observed that e-speed start is able to increase throughput between 700-51100% and network link utilisation by 37.5% as against the convectional slow-start technique in TCP.

I. Introduction

TCP New Reno was upgraded from experimental status to full protocol status in 2004. Several proposals and researches had been put forward to improve its performance. Many of these proposals tend to improve TCP in a high speed network where it has been shown that TCP mechanism may lead to network resources underutilization [14]. TCP Westwood proposed bandwidth estimation as congestion measure [10], it specified that a TCP sender continuously computes the connection bandwidth estimate by properly averaging the returning ACKs and the rate at which the ACKs are received. After a loss has occurred, the sender uses the estimated bandwidth to properly set the sending rate and the congestion window. This is an improvement on standard TCP which half its window on loss detection [10]. However, TCP Westwood have not proved better in term of stability and fairness when it co-habit with the standard TCP and its suitability for general deployment has not been ascertained.

Other proposed protocols in this category include XCP [5] which requires modification to router algorithm. However, it is not visible to modify all existing routers' algorithms hence XCP will remain experimental protocol for a long time. HSTCP [3] was also designed for high bandwidth delay product. It uses loss-delay to detect congestion. Wei in [16] proposed the FAST TCP which uses delay instead of loss to signal congestion. Other protocols in this category include HCTP [8], BIC TCP [17], STCP [6], CUBIC TCP [17] etc. These protocols modify the window growth function of TCP to match large bandwidth delay product. It appears easy to accomplish this modification but the issue of fairness with these protocols has remained a challenge. Fairness in this case involves both intra and inter-protocol fairness. In addition, none of these protocols modified the startup behaviour –slow start of AIMD, which has come under investigation lately.

There have been several modifications of the slow start stage to overcome the problem of performance associated with it. According to Liu in [9], there are three lines of studies to slow start mechanism modification. The first approach uses capacity estimation techniques to estimate available bandwidth and sets the congestion window size using the estimated bandwidth. Partridge C et al in [12] proposed Swift Start for TCP. Swift start employs an initial window (cwnd) of 4 packets thereafter estimates the available bandwidth in the first round trip time using packet pair. Swift start uses the estimated bandwidth to compute the bandwidth delay product (BDP) of the network and set cwnd to a percentage of the estimated BDP. Lawas-Grodek and Tran in [7] carried out a performance evaluation of Swift Start and submitted that Swift Start improves network performance when the network is not congested. However, when the network becomes overflowed the estimation of the cwnd drifts away from accuracy. This drifting is due to retransmission of delayed or lost ACKs and Retransmission Time Out (RTO) timeouts.

A second approach employs sharing of congestion information among connections in the network. An example of congestion control algorithm that used this approach is the Congestion manager proposed by [2]. Congestion manager collects congestion status information and feedback from receivers, shares the information with endpoints and connections in the network which enables connections determine the congestion status of the network and thereby determine an initial sending rate. The congestion manager has a weakness of being beneficial to only connections that are initiated almost at the same time. Secondly, it is only those connections and endpoints that supplied feedback that can benefit from this scheme. And lastly, there is an approach that

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_85,

requires explicit network assistance to determine a practicable starting sending rate called Quick Start [4].

According to [4], the experimental Quick start TCP extension is currently the only specified TCP extension that realizes a fast startup. A large amount of work has already been done to address the issue of choosing the initial congestion window for TCP. RFC 3390 [1] allows an initial window larger than one packet. Quick start is based on the fact that explicit feedback from all routers along the path is required to be able to use an initial window larger than those specified by RFC 3390.

In quick start proposals, a sender (TCP host) would indicate its intention to send at a particular rate in bytes per second. Each router along the path could approve, reduce, disapprove or ignore the request. Approval of the request by the router indicates that it is being underutilized currently and it can accommodate the sender's requested sending rate. The quick start mechanism can detect if there are routers in the path that disapproved or do not understand the quick start request. The receiver communicates the response to the sender in an answering TCP packet. If the quick start request is approved by all routers along the path, then the sender begins transmission at the approved rate. Subsequently, transmissions are governed by the default TCP congestion control mechanism. If the request is not approved, then the sender transmits at the normal TCP startup speed [15].

According to [15], TCP is effective for both flow control and congestion control. TCP flow control is a receiver-driven mechanism that informs the sender about the available receive-buffer space and limits the maximum amount of outstanding data. Flow control and congestion control are independent while size of receive buffer space depends on the capacity of the receiver network. However, if the TCP is used with a link with large bandwidth-delay product, both congestion window and flow control window need to be large in other for TCP to perform well [4]. This makes both flow control and congestion control to overlap. The quick start TCP extension assumes independence of flow and congestion control which is not true and makes it not appropriate for the global internet.

Another issue with the quick start approach is that the starting speed of a connection using quick start may send thousands of packets in one RTT. The implication of this is that, thousands of packets may be dropped in one RTT, thereby necessitating a retransmission of these packets which may in turn be dropped and retransmitted over and over again, thereby saturating the network which can lead to congestion. In addition, most routers and other network components on the global internet are not built with capabilities to be quick start aware. The implication is that they need either to be replaced or modified before deployment. Hence, we propose a milder approach to fast startup which called e-speed start. This is a form of fast startup that does not require routers' response or modification. Rather, it employs end-to-end principles to determine a suitable startup speed. This proposal is adaptable to the current internet as well as the future gigabit internet.

The remaining of this paper is organized as follows: section II presents the existing TCP slow start mechanism. Section III introduces our proposed approach called e-speed start, the design of e-speed start and its mechanism. Section IV discusses initial results obtained. Section V concludes the paper.

II THE EXISTING MODEL

Window based congestion control has four stages; this is to say that the life time of a connection passes through possibly four stages before connection termination. These are the slow start, congestion avoidance, fast retransmit and fast recovery. Most research efforts in the past have been concentrated on the congestion avoidance stage and several advances have been made until recently when research efforts began to focus on the slow start stage due to advances in network speed technologies which made it possible to have gigabit networks operating at a speed of 1 gigabit/second (gbs) and above, compared to the maximum speed of 100 megabits per second (mbps) that existing networks transmit. The implication of this higher speed of transmission is that the current congestion control techniques being used on the internet may not be suitable for the best effort network in the future. The slow start stage is modeled as described below. The algorithm below applies:

The Slow Start Algorithm

1. Start with a window size of 1 i.e cwnd=1
2. Increase the window size by 1 i.e. cwnd=cwnd+1; for every ACK[1] received. Repeat step 2 until the ssthresh is reached OR Packet loss is detected
4. If ssthresh is reached, go to congestion avoidance phase.
5. Else If a packet loss is detected;
   set to ssthresh[2] = 0.5*cwnd[3]
   set to cwnd =1
   go to step 2
6. Endif

Note: Set the initial value of ssthresh to a fraction of the maximum window size. This is determined at the beginning of transmission. Table I shows the simulation of the slow start algorithm stated above. From table I the window size is expressed as

$$W(nT + m/c) = 2^{n-1} + m + 1, \; 0 \le m \le 2^n - 1 \qquad (1)$$

And the maximum window size in a cycle is derived as

$$W(m) = 2^n \qquad (2)$$

There will be two slow start phase. The first phase, buffer overflows and there is a packet loss which reduces the window size to 1 and slow start is re-entered. Equations 3 and 4 describe the ssthresh in both phases respectively.

$$ssthresh = (cT + B)/2 \qquad (3)$$

$$ssthresh = Min\,(2B-1, (cT + B)/4) \qquad (4)$$

where B is buffer size of link, c is the link capacity, T is sum of propagation and queuing delay, W(m) is maximum window size in a cycle, n is the number of cycles and m is number of

[1] Acknowledgment
[2] Slow start threshold
[3] Congestion window

TABLE I
MODELLING THE SLOW START

Cycle[1]	Time	Acked packet	Window size	Max packet released
Cycle 0	O	-	1	1
Cycle 1	T	1	2	3
Cycle 2	2T 2T + 1/c	2 3	3 4	5 7
Cycle 3	3T 3T + 1/c 3T + 2/c 3T + 3/c	4 5 6 7	5 6 7 8	9 11 13 15
Cycle 4	4T 4T + 1/c 4T + 2/c 4T + 3/c 4T + 4/c 4T + 5/c 4T + 6/c 4T + 7/c	8 9 10 11 12 13 14 15	9 10 11 12 13 14 15 16	17 19 21 23 25 27 29 31
Cycle 5	5T 5T + 1/c 5T + 2/c …	16 17 18	17 18 19	33 35 37

RTT within a cycle. ((4) follows from the fact that ssthresh is half of window size.) (1) to (4) was modified from [18].

III. WORK DONE

In our approach presented here, we attempt to find a balance between slow start low startup speed transmission and the large number of segment transmission of quick start. The approach introduces two network environmental variables hose, periodic values are determined and assigned. These values are based on two operating conditions prevalent in the network, peak or off-peak network utilization periods. Network utilization periods are representations of network users' behaviour, users are usually not using the network at certain times e.g. night session, holiday period etc. Network utilization is used to determine whether or not to employ conventional slow-start mechanism for transmission. At a particular point in time, if the network utilization is off peak the available bandwidth[20] module is used to estimate the available bandwidth of the network. This value is used to compute a factor β which in turn is used to determine an appropriate initial congestion window size and hence the initial start-up speed. On the other hand, if the available bandwidth is too low, the connection defaults to slow start. At any sign of congestion indication within the network, the e-speed start mechanism is dropped and transmission proceeds via the conventional TCP.

Approaches to congestion control are of two types; rate based control and window based control. A window based congestion control approach is presented in this work. A window based congestion control uses a congestion window to determine the sending rate of the end nodes. Here the concept of end-to-end control becomes useful. End-to-end control builds congestion control complexity at the end nodes while the link is assumed unintelligent and less complex.

A. Design of the E-Speed Start

The current design of the slow start state of the internet as expressed in New Reno [19] and other proposed modification specified a single start speed for the TCP. Although, slow start can work fairly well with networks with low bandwidth delay product, its performance suffers from severe inefficiency when it is applied to high bandwidth delay product (BDP) networks. This is so because it takes too long for the sender to reach an available high rate.

The slow start had been the traditional determinant of start up speed for all proposed window based congestion control algorithm until recently when [4] proposed the quick start modification to TCP slow start algorithm. With the quick start proposal, large data could be sent using an initial large window that sends thousands of packets at the startup stage. However, this has a requirement that the network presents high bandwidth capacity to be able to transmit at the high rate of quick start. This also implies that thousands of packets may be lost in a single round trip time thereby overburdening the network.

Our proposal is to build an algorithm that is intelligent enough to be aware of the availability or non availability of certain environmental factors that determine an initial speed for the startup stage of the window based congestion control algorithm. In this work, it is proposed that initial startup speed may be chosen from two options (e-speed start and slow start). However it is possible to make startup speed selectable from an n-arry set of algorithms instead of binary option as proposed in this work.

B. Environmental Factor

The environmental factor referred to in this work performs two functions along with a bandwidth estimation function that determines the available bandwidth. These are the time function and user input function. The time function obtains the system's time of the day and expresses it as 'peak' or 'off peak' time. An initial definition of peak and off-peak period may be set and may be refined or redefined by using network usage statistics. User inputs specify whether the day is a night period, workday or holiday. Some networks are not heavily used in some day of the week, for instance, over the weekends or in the night. These periods may afford a higher startup speed without necessarily congesting the network and thereby optimizing the use of network resources – bandwidth.

The e-speed start is designed to be an end-to-end control not requiring any input or modification from the network unlike the quick start that require routers input/modification. The design of the E-speed start decouples speed determination from actual data transfer at this stage. Hence the startup stage is divided into two: speed determination stage and the data transfer stage.

C. Speed Determination Stage

At this stage, two main environmental variables are introduced and their values determined in the following order:

- Time-of-day: a binary value of peak or off-peak
- User-input: a binary value of workday or holiday.

In standard TCP, after handshake, the connection transits from idle to slow start state where data transmission proceeds immediately. In e-speed start, successful handshake transits into the speed determination state. At the speed determination state, the system time is captured into the variable Time-of-day and the value is compared with a table that contains data collected from network usage statistics. If bandwidth usage is high at this time (as contained in the network usage statistics table), the value of Time-of-day is set to peak and if bandwidth usage is low, it is set to off-peak. It is assumed that if network usage is low for a period of time at a particular time of the day, it is likely to be low always at such period of time and vice versa. However, the determination of the period of the day to be categorized as low or high bandwidth usage time is not trivial. We intend to use fuzzy logic to determine the probability of low or high bandwidth usage time. If the value of Time-of-day is peak the connection activates user input module and if it is off-peak it proceeds to bandwidth estimation.

The second environmental variable is user-input, the user optionally responds to the prompt, "is it an holiday?" for example, during connection attempt. If yes, the speed determination module default to bandwidth estimation to determine available bandwidth, otherwise it proceeds to slow start. Time-of-day and User-input are two variables that determine if a connection proceeds to bandwidth estimation state or directly to slow start state.

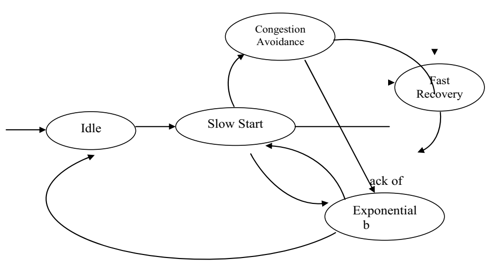


Fig 1: TCP state transition using Slow Start

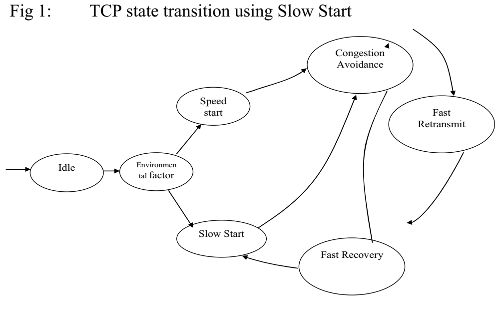


Fig 2: TCP state transition using e-speed start

The system transits into bandwidth estimation mode if the value of Time-of-day variable is off-peak or if user-input is yes in response to the holiday prompt. The bandwidth estimation module estimates the available unused bandwidth. If bandwidth usage is low and the estimated available bandwidth is high, the system enters e-speed start otherwise slow start is activated.

From figure 3 below, the time function captures the time of connection, determines if it can be categorized as off-peak or peak period, and passes control to bandwidth estimation function or user input as the case may be. The bandwidth estimation function uses the packet pair algorithm to determine the available bandwidth and thereby passes control to either the slow start or e-speed start. If control is passed to the slow start, it proceeds as in conventional TCP. If the user-input module is activated, the user is prompted with an optional message e.g. "press spacebar if it is an holiday". If user responds to this message the bandwidth estimation module is activated else the slow start state is entered into and data transmission starts. If the bandwidth estimation function returns a high value for available bandwidth, e-speed start can be used for startup for data transmission.

The e-speed Start Algorithm

0. Estimate the value of β using available bandwidth. If β=1 go to 1else if β>1 go to 7
1. Start with a window size of 1 i.e cwnd=1
2. Increase the window size by 1 i.e. cwnd=cwnd+1; for every ACK received. Repeat step 2 until the ssthresh is reached OR Packet loss is detected
4. If ssthresh is reached, go to congestion avoidance phase.
5. Else If a packet loss is detected;
   set to ssthresh = 0.5*cwnd
   set to cwnd =1
   go to step 2
6. Endif
7. If β=n, Increase the window size by 2^β i.e. cwnd=cwnd+2^β; for every ACK received. Repeat step 7 until ssthresh is reached OR Packet loss is detected
8. If ssthresh is reached, go to congestion avoidance phase.
9. Else if a packet loss is detected;
   set to ssthresh = 0.5*cwnd
   set to cwnd =1
   go to step 2
10. Endif

Using decision table to describe these scenarios, we obtain table II below: From table II, it is observed that E-speed is used for startup at 37.5% of the time while slow start still dominate most of the connections at 62.5% of the time. This implies that the link/network would have been underutilized 37.5% of the time. It is clear that e-speed start optimizes the available bandwidth utilisation at these times for faster connection and data transfer.

IV DISCUSSION OF E-SPEED START MECHANISM.

In e-speed start mechanism, new connection starts by TCP sending an initial packet of 1. Thereafter, the congestion window is grown by (5);

$$cwnd = 2^{n\beta} \quad (5)$$

This is a modification of maximum window size per cycle from (2) above where n is the number of RTT and β is a function of the available bandwidth. For instance, if β is set to two, then four packets are sent per each ACK received and it continues until there is a notification of congestion. The resulting table as compared with the conventional slow start is presented table III and IV. From table III, it is observed that when β=1, the throughput at RTT=7 is 128 packets while β=2 yielded over 16000 packets for the same number of RTT. This represents 12400% increase at slow start stage as long as other operating conditions remain stable.

For case of β=2, the new table for packet transmission is as presented in table V: The model describing e-speed start, window size is given by (6)

$$W(nT + m/c) = 2^{\beta(n-1)} + 3m + 3, 0 \leq m \leq 2^{\beta(n-1)} - 1 \quad (6)$$

and maximum window size in a cycle

$$W(m) = 2^{\beta n} \quad (7)$$

Equations 6 and 7 form the bases for the window growth function in the e-speed start..

From table III, when β=2, four packets are transmitted in response to one ACK received. This yielded a higher throughput. For instance a total of 4,096 packets are sent by e-speed start as against 64 packets in the slow start at RTT=6. For β=3, the same RTT yields 262,144 packets. Figure IV shows that by the second cycle, e-speed start speed can increase TCP slow-start throughput between 700 and 51100%. It is important to emphasise here that network capacity utilisation is the main trust of the e-speed start algorithm. The algorithm can adapt its speed from network to network hence it is both backward compatible with networks with low bandwidth delay product (BDP) as well as and forward compatible with networks with higher BDP

V. CONCLUSION

We have presented here, a modification of the slow start stage of the window based congestion control algorithm. This modification is different from other modifications in literature as it achieved the following. It clearly decoupled speed determination from data transfer mode at the slow start. This gives room for a differential start up speed depending on the environmental factors like peak or off-peak period, a weekend, night session etc. In addition, available bandwidth is major determination of the startup speed. It is used to determine a factor β which is used as a multiplier of n, the power in 2^n. If the calculated value of β is 1, then TCP proceed with the normal slow start. If β is 2, TCP starts with a transmission of one packet and transmit 4 packets for each ACK received, If β is 3, TCP starts with a transmission of 1 packet and transmit 8 packets for each ACK received. β can take any integer value provided the available bandwidth can support the resulting transmission speed. In addition, we propose that TCP start-up may be selected from n-arry possible algorithms instead of a single slow start as in the New Reno. We have proposed a binary selection for start up stage and suggested a further investigation to using an n-arry selection mechanism for AIMD. This work is ongoing and these functions will be analyzed for stability, e-speed start will be benchmarked against major window based congestion control algorithms to determine its fairness in terms of intra protocol and inter protocol fairness.

TABLE II : DECISION TABLE FOR E-SPEED START

Conditions	Rules															
Time of Day is off-peak	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N
Available bandwidth is high	Y	Y	Y	Y	N	N	N	N	Y	Y	Y	Y	N	N	N	N
User Input is yes	Y	Y	N	N	Y	Y	N	N	Y	Y	N	N	Y	Y	N	N
It is an holiday	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Action	Result															
Go to E-speed start	X	X	X	X					X	X						
Go to slow start					X	X	X	X			X	X	X	X	X	X

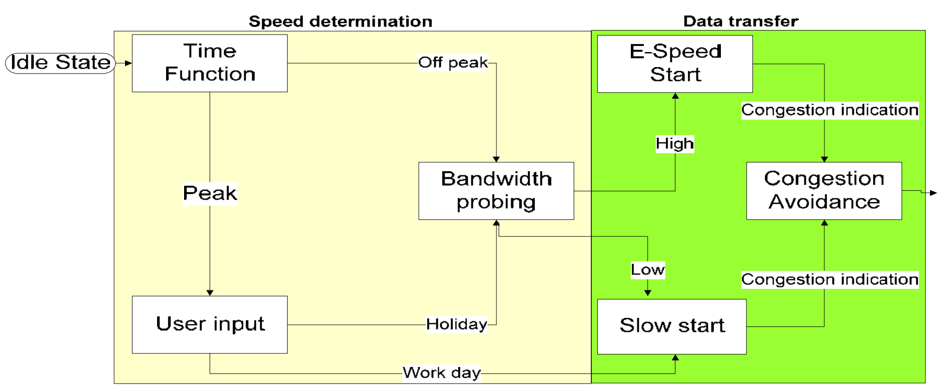


Fig 3: Schematic diagram for E-Speed Start

TABLE III
E-SPEED START TRANSITION CYCLES WITH β =2 and 3

Cycle	Congestion Window for Slow-Start (β=1)	Congestion Window for E-speed Start (β=2)	Congestion Window for E-speed Start (β=3)
0	1	1	1
1	2	4	8
2	4	16	64
3	8	64	512
4	16	256	4,096
5	32	1,024	32,768
6	64	4,096	262,144
7	128	16,384	2,097,152
8	256	65,536	16,777,216
9	512	262,144	134,217,728
10	1,024	1,048,576	1,073,741,824
11	2,048	4,194,304	8,589,934,592
12	4,096	16,777,216	68,719,476,736

TABLE V
MODELING THE E-SPEED START WITH β=2

Cycle	Time	Acked packet	Window size	Max packet released
Cycle 0	O	-	1	1
Cycle 1	T	1	4	5
Cycle 2	2T	2	7	9
	2T + 1/c	3	10	13
	2T + 2/c	4	13	17
	2T + 3/c	5	16	21
Cycle 3	3T	6	19	25
	3T + 1/c	7	22	29
	3T + 2/c	8	25	33
	3T + 3/c	9	28	37
	3T + 4/c	10	31	41
	3T + 5/c	11	34	45
	3T + 6/c	12	37	49
	3T + 7/c	13	40	53
	3T + 8/c	14	43	57
	3T + 9/c	15	46	61
	3T + 10/c	16	49	65
	3T + 11/c	17	52	69
	3T + 12/c	18	55	73
	3T + 13/c	19	58	77
	3T + 14/c	20	61	81
	3T + 15/c	21	64	85
Cycle 4	4T	22	67	89
	4T + 1/c	23	68	93
	4T + 2/c	24	71	97
	4T + 3/c	25	74	101
	4T + 4/c	26	77	105
	4T + 5/c	27	80	109
	4T + 6/c	28	83	113
	…			

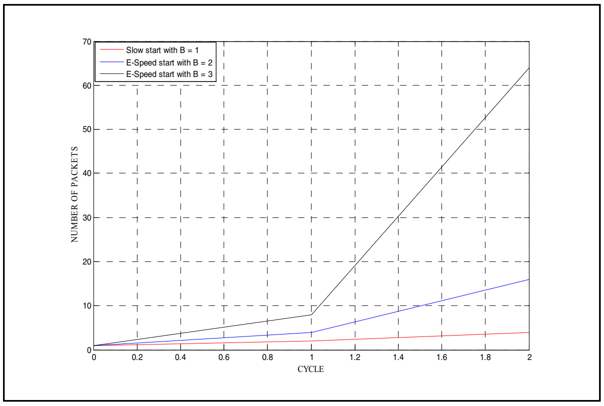


Figure. 4: Graph of slow start and e-speed start with β=2 and β=3

REFERENCES

[1] M. Allman, Increasing TCP's initial Window, RFC3390, 2002 www.faqs.org/rfcs/rfc3390.html.

[2] H. Balakrishman., H. Rahul. and Seshan, An integrated .congestion management architecture for Internet hosts. In Proceedings of ACM SIGCOMM '99, Cambridge, MA. 1999.

[3] S. Floyd High-speed TCP for large congestion windows RFC 3649, December 2003.

[4] S Floyd., M. Allman, A. Jain, P. Sarolahti. Quick-Start for TCP and IP. RFC 4782. 2007.

[5] D. Katabi, , M .Handley and C. Rohrs. Congestion control for high bandwidth-delay product networks. In proceedings on ACM Sigcomm. 2002.

[6] T. Kelly, Scalable TCP improving performance in high speed wide area networks. Computer communications Review 2003. 32(2).

[7] F. Lawas-Grodek and D. Tran. Evaluation of Swift Start TCP in Long Delay Environment. Technical report, Glenn research Center, Cleveland, OH. 2004.

[8] D. Leith, R. Shorten and Y. Li. H-TCP: A framework for congestion control in high speed and long distance networks. Hi tech report. 2005. http://www.hamilton.ie/net/htcp.

[9] D. Liu., M. Allman., S..Jin, and L. Wang. Congestion control without a startup phase. In 5th International Workshop on Protocols for fast long distance networks, Marina Del Rey, CA, 2007.

[10] S. Mascolo, C. Casetti., M. Gerla., M. Sanadidi., R. Wang. TCP Westwood: end-to-end bandwidth estimation of efficient transport over wired and wireless networks, in: Proceedings of the ACM Mobicom, Rome, Italy. 2001.

[11] A. .Oluwatope,, A. Obadire,, G. Aderoumu, and M. Adigun, End-to-End evaluation of selected TCP variants across a hybrid wireless network, Issues in Information Science and Technology, 2006.Vol. 3.

[12] C. Partridge, D. Rockwell, M. Allman, R. Krishnan, and J. Sterbenz, A swifter start for TCP. In Technical Report TR-8339, BBN Technologies. 2002.

[13] I. Rhee, and L. Xu. CUBIC: A new TCP – friendly high speed CTP variant. In International Workshop on protocols for fast Long – distance Networks, Lyon. 2005.

[14] P. Sarolahti, M. Allman, and S. Floyd. Determining an Appropriate Sending Rate Over an Underutilized Network Path. February. 2006. http://www.icir.org/floyd/quickstart.html

[15] M. Scharf, and S. Floyd, TCP Flow Control for Fast Startup Scheme draft-scharf-tcpm-flow-control-quick-start-00.txt

[16] D. Wei, C. Jin, S. Low, and S. Hedges, Fast TCP: motivation, architecture, algorithms, performance IEEE-ACM Transaction on networking, 2006. 14(6): 1246 – 1259.

[17] L. Xu, K. Harfoush, and I. Rhee,. Binary Increase Congestion control for fast long distance networks. In proceedings of IEEE INFOCOMM '04. 2004

[18] R. Srikant. "The Mathematics of Internet Congestion Control." Birkhauser, 2004. Amazon Link. Notes and Errata

[19] S Floyd., T. Henderson and A. Gurtov. The New Reno modification to TCP's fast recovery algorithm RFC 3782 (2004).

[20] A. O. Oluwatope, G. A. Aderounmu, E. R. Adagundo, O.O. Abiona and F.J. Ogwu. Available Bandwidth Based Congestion Avoidance Scheme for TCP: Modelling and Simulation. Advances in Computer and information Sciences and Engineering, Springer Science +Business Media B.V. 2008. Pp 85-89

Common-Wavelengths Design Scheme for Efficient Management of Spare Resources in WDM Networks

Ardian Greca
Department of Computer Sciences, Georgia Southern University
P.O. Box 7997 Statesboro, GA, naidrag@ieee.org

***Abstract*-The author addresses the problem of self-healing capability to recover from network failures for WDM optical networks. One problem in realizing such survivable WDM networks is the inefficient utilization of large spare resources (wavelengths) designed for restoration of the failed traffic, which might not happen frequently. We study "common-wavelengths" design scheme to increase the utilization efficiency of these resources. The "common-wavelength" is an approach that allows some working paths under the failure to efficiently use the available common-wavelengths in the network. The effectiveness of this proposed approach and the effect of physical and traffic connectivity are investigated via simulations. Thus, the survivable WDM networks can be cost effectively designed according to the reliability levels of all working paths in the network.**

I. Introduction

The survivability of a network against unexpected failures is one of the important characteristics for optical WDM networks, due to a huge traffic loss. Network survivability can be carried out by using self-healing techniques, which can restore the failed traffic automatically by rerouting the failed paths to other unaffected alternate routes when the failure occurs. In order for the self-healing technique to succeed a sufficient amount of spare resources (wavelengths) should be pre-designed to complete the restoration under the failure scenario. While the amount of spare resources depends on several factors such as type of failure considered, network topology, self-healing scheme, still a large amount, exceeding 40% of working resources, might be required for even simple failure scenarios. In order to realize a cost-effective survivable network the amount of spare resources in the network should be decreased. Many researchers have studied the cost-effective ways to reach the desired level of availability and enhance the survivability performance of the network [1-3 and the references therein]. In [1] authors studied how the optical layer survivability can be introduced in SDH over WDM transport networks. Researchers in [2] studied the survivable traffic grooming in WDM mesh networks. They proposed the protection-at-lightpath (PAL) level and the protection-at-connection (PAC) level, and concluded that PAL outperforms PAC for the dynamic provisioning shared-mesh-protected sub-wavelength connections against single-fiber failures. In [3] authors proposed a cost-effective provisioning approach to provide differentiated per-connection-based services to carry connections by appropriate routes and to protect them according to their availability requirements.

Furthermore, considerable research efforts have been dedicated to the study of survivability mechanism in WDM and IP/GMPLS network respectively [5-15]. Some of these efforts have been put on the WDM layer, some others on the IP/GMPLS layer, and recently many others have addressed the multilayer survivability between these two layers. Although there is a significant progress done so far, there is still a need for research focused on the WDM layer survivability mechanisms. In order to provide a comprehensive and efficient survivable network service, let us compare the pros and cons for each of these layers. The survivability mechanisms in WDM layer are faster and more scalable than those of IP/GMPLS layer; however, they cannot handle faults such as router fault and service degradation that usually are covered by IP/GMPLS survivability techniques. On the other hand, survivability mechanisms at the IP/GMPLS layer are usually slower and less scalable. One such example is a link failure which can result in a failure of thousands IP layer traffic. Thus, providing a better survivability in WDM layer can improve the over all performance of the network and might even lead to efficient ways for multilayer survivability.

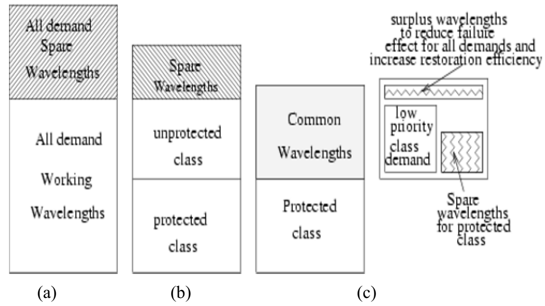


Fig. 1. Failure restoration designs.

Researchers at [16] have described the concept of common pool survivability, which allows some spare capacity of the lower SDH layer to be used for recovery of the upper ATM layer traffic. The approach is studied in the framework of multilayer survivability and can perform poorly if it is not carefully designed

In this paper we propose and study the "common-wavelengths" design scheme at the WDM layer to increase the efficiency of resource utilization in the network under failure. The approach considers "multi-reliability class" concept to design the "common-wavelengths" resources needed for protection of the affected traffic and to reduce the number of spare wavelengths in the network. The method uses integer linear programming (ILP) model. In our simulations we show that the scheme reduces spare resources in the network. To the best of our knowledge, we are not aware of such an approach being studied so far for this layer.

The rest of the paper is organized as follows. Section 2 explains the concept of "common wavelengths" and the policies used to realize

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_86, © Springer Science+Business Media B.V. 2010

such a design. Section 3 discusses the formulation of the problem using ILP. Simulated network topology and simulation results are discussed in Section 4. Finally, conclusions are given in Section 5.

II. The concept of common-wavelengths

Figure 1a, shows the straight forward approach where there is no reliability class among traffic demands. This leads to the most expensive approach since spare resources have to be designed for all demands. A more efficient approach is shown in Fig. 1b, where demands are grouped in two classes, protected class and unprotected one. In this approach the total spare resources can be decreased since they will be designed only for the demands of the protected class. The best approach appears to be if working resources of both the unprotected class and spare resources are combined together to assure the recovery under the failure. These resources are referred to as "common-wavelengths". The policy of using "common-wavelengths" is as follows:

1. During the normal operation of the network, "common-wavelengths" will be used by demand of low priority class.
2. Under the failure scenario, alternative paths for protected paths will have the first priority to use these "common-wavelengths". If the available resources are not enough some of the demands of the unprotected class have to release their wavelengths for the alternative paths.
3. Low priority class paths can either be randomly dropped or their bandwidth requirement can be decreased by average amount.

This approach reduces the amount of spare resources in the network and leads to a cost-effective solution for WDM networks. The drawback of this scheme is that it will degrade the QoS for some of the demands in the unprotected class. But this degradation, if it happens, will be only during the failure time which will not occur frequently.

III. Problem Formulation

This study uses the same approach as in [4] and considers static optical path assignment; that is, a request for setting up a set of optical paths is given first. These optical paths are not released once they are set up. The optical paths are assumed to be wavelength paths, (WP), (lightpaths). Thus, the wavelength of a path is not converted at any transit node in the network. We consider an optical network that offers *k* wavelengths for setting up WPs. Let $G(V,E)$ denote the directed graph that models the network. This model is equivalent to the model with *k* separate networks, each of which has the same topology as $G(V,E)$ and provides only one wavelength (Fig. 2). Let the graphs of these *k* single networks be $G^{(1)},...,G^{(k)}$. Also, let $v^{(i)}$ denote the vertex in $G^{(i)}$ corresponding to the vertex $v \in V$ in the original graph. It is possible to combine k $G^{(i)}$ graphs into one extended graph in the following way. Add vertices $s_1,\dots s_{|v|}$ and create an edge from s_v to $v^{(i)}$ for $|V|$ every $v \in V$ and every $1 \leq i \leq k$. Similarly, add $|V|$ more vertices $t_1, \dots t_{|v|}$ and create an edge from $v^{(i)}$ to t_v for every $v \in V$ and every $1 \leq i \leq k$. With this graph conversion, the routing and wavelength assignment problem becomes a multi-commodity flow problem for each sub-graph $G^{(i)}$. Let $D = \{d_1, \dots d_j\}$

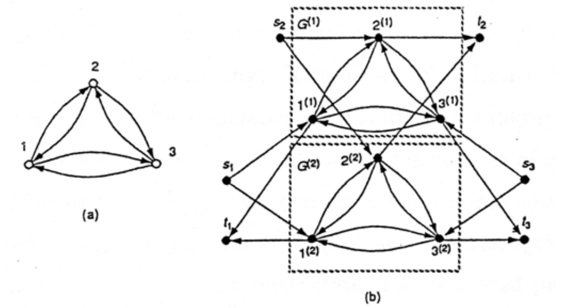


Fig. 2. Original and expanded graph of a network.

be the set of optical paths demanded, and let *a(d)* and *b(d)* denote the source and the destination of the path $d \in D$ respectively. Suppose that a flow of the commodity associated with the path *d* is directed from $s_{a(d)}$ to $t_{b(d)}$ in the converted graph, and the flow volume of each commodity is 1. If the capacity of every edge is 1 and if the flow volume on an edge is restricted to 0 or 1, the route of the flow will have the following characteristics:

- The flow of a commodity passes a single route.
- The flow of a commodity goes through only one of the subgraphs $G^{(1)}, ..., G^{(k)}$.
- At most one commodity can go through each edge of G.

Among these characteristics, the first and the third one follow straightforwardly from the edge capacity and the integer restriction. The second characteristic comes from the fact that in order to satisfy the wavelength continuity constraint of the WP there are no routes between subgraphs.

In summary we have the following notation:

- V the set of nodes in the network.
- E the set of links in the network.
- Λ the set of wavelengths in the network (subgraphs).
- G(V,E) network graph.
- $G^{(i)}(V,E)$ network graph for each wavelength.
- $G^{gen}(V_1,E_1)$ modified network graph with network topology, links in each subgraph, wavelengths and added nodes.
- l_{in} link that enters a node.
- l_{out} link that exits a node.

We consider variable y_{max} to denote the maximum number of demands going through each link of the general graph. Also let x_{dl} be a variable that denotes a demand d going through a link l of the general graph. x_{dl} =1 if demand d goes through the link l and 0 otherwise. The formulation of the problem is as follows:

Minimize y_{max}
Subject to:
1. the sum of all demands on each link should be less or equal than y_{max}

$$\sum_{d=1}^{D} x_{dl} \leq y_{\max} \quad \forall l \in E_1, \lambda \in \Lambda$$

2. all paths that share a link should have different wavelengths

$$\sum_{d=1}^{D} x_{dl} \le 1 \qquad \forall l \in E, \lambda \in \Lambda$$

3. each demand uses the same wavelength through its path

$$\sum_{l=1}^{E_1} x_{dl} \le 1 \qquad \forall d \in D, \lambda \in \Lambda$$

4. There are no loops

$$\sum_{l=1}^{E_1}\sum_{d=1}^{D} x_{dl_{in}} + \sum_{l=1}^{E_1}\sum_{d=1}^{D} x_{dl_{out}} \le 2 \qquad \forall n \in V, \lambda \in \Lambda$$

5. Conservation law

$$\sum_{l=1}^{E_1}\sum_{d=1}^{D} x_{dl_{in}} - \sum_{l=1}^{E_1}\sum_{d=1}^{D} x_{dl_{out}} = \begin{cases} 1 & \text{if } n \text{ is the origin} \\ -1 & \text{if } n \text{ is the destination} \\ 0 & \text{otherwise} \end{cases}$$

6. All variables can be either 1 or 0

$$x_{dl} = 1 \, or \, 0 \;\; \forall d \in D, \lambda \in \Lambda$$

IV. Simulations And Results

The network models used for our simulations are shown in Fig. 3. The parameters of these networks are the number of network nodes and the physical connectivity. The number of fiber links between all network nodes defines the physical connectivity in our simulations. The total number of fibers per link is considered 1. Then we generate 4 different traffic patterns (demands) randomly chosen for the network to be fully loaded. Each of these traffic patterns is a combination of three classes of service. The first class consists of 20%, the second class consists of 30% and the third class consists of remaining 50% of the total demand. Each is chosen randomly. In case of a failure the first and second class will have priority for restoration and the third one would have to release some of its paths to make the "common-wavelengths" available to the traffic demand. Based on the requirements described above the ILP model is build to solve the problem as follows:

1. Calculate the number of wavelengths needed to satisfy the demand in the network.
2. Generate a random link failure and calculate the number of wavelength needed to satisfy all traffic in the network.
3. Calculate the spare wavelengths needed to satisfy the affected wavelength paths.
4. Release some or all traffic from the third class and calculate the number of wavelengths needed to satisfy the affected traffic and released traffic.
5. Repeat the above steps for different failure scenarios and different traffic patterns.

Results of simulations for the three schemes described in Fig. 1 considering 15 nodes α=0.238 network are shown in Fig. 4. We can see that on average scheme (a) requires about 40%, scheme (b) requires about 29% and our "common-wavelength" scheme (c) requires only 10% of total resources. The same trend is observed for the other networks presented in Fig. 3a (not presented here).

We determine the connectivity of the traffic matrix from the number of communication paths in it and the number of nodes. The traffic matrix, like physical connectivity itself, is represented by a graph. The difference between the traffic matrix and the physical network is that the edges in the graph are represent the end-to-end connections. Then we studied the impact of the traffic connectivity on the number of wavelengths required to fulfil the given demand. Two regular networks, consisting of 9 nodes α=0.5 and 15 nodes α=0.238 respectively, were considered. Results are shown in Fig 5. As we observe from this figure the increase in the traffic matrix will result on the increase of the number of required wavelengths. Furthermore, a decrease on the physical connectivity will result on the increase of the

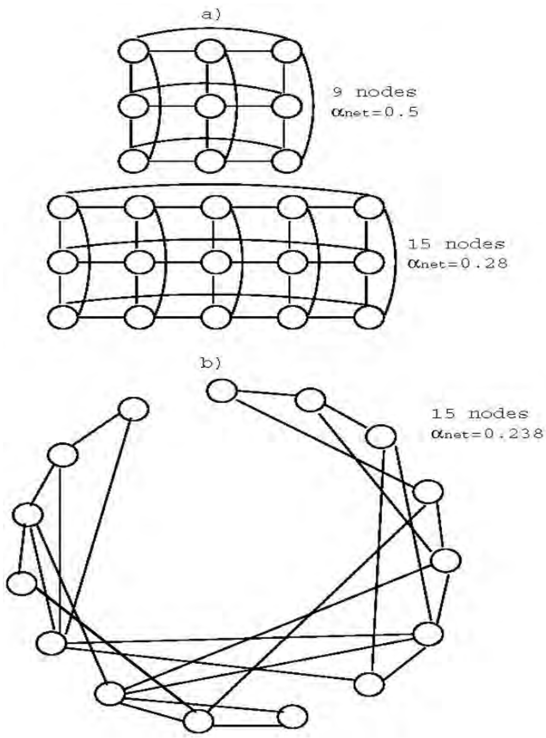


Fig. 3. Topologies of the photonic networks under investigation: a) regular networks and b) network with random topology.

Fig. 4. Simulation results for 15 nodes α=0.238 network.

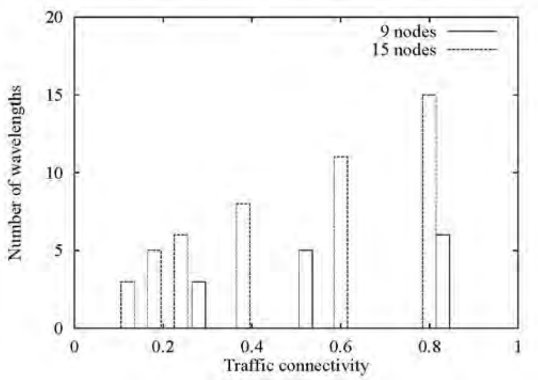


Fig. 5. Wavelength requirement vs. traffic connectivity in regular network topologies.

number of wavelengths. This increase is almost linear. These results are in line with other observations given in literature.

V. Conclusions

In this paper we studied the design of "common-wavelengths" scheme which can be used to build cost-effective survivable WDM networks. We used ILP to solve the problem. Simulation results confirmed that the proposed scheme reduces the number of required wavelengths when a failure happens in the network.

References

[1] P. Demeester, M. Gryseels, K. van Doorselaere, A. Autenrieth, C. Brianza, G. Signorelli, et al. "Network resilience strategies in SDH/WDM multilayer networks", 24th European Conference on Optical communication, Vol. 1, pp. 579-580, September 1998.

[2] C. Ou, K.Y. Zhu, B. Mukherjee, "Survivable traffic grooming in WDM mesh networks", IEEE OFC, Vol. 2, pp. 624-625, March 2003.

[3] J. Zhang, et al IEEE OFC, Vol. 2, pp. 622- 623, March 2003.

[4] A. Greca, et al IEEE Workshop on HPRS, pp. 146-149, Dallas 2001.

[5] S. Ramamurthy, B. Mukherjee, "Survivable WDM mesh networks, part I – protection", Proc. IEEE Infocom 99.

[6] S. Ramamurthy, B. Mukherjee, "Survivable WDM mesh networks, part I – restoration", Proc. ICC 99.

[7] O. Gerstel, R. Ramaswami, "Optical layer survivability – an implementation perspective", IEEE Journal of Selected Areas in Communications, Vol. 18, No. 10, pp. 1885 – 1889, October 2000.

[8] J. Manchester, P. Bonenfant, C. Newton, "The evolution of transport network survivability", IEEE Communication Magazine, August 1999.

[9] O. Gerstel, R. Ramaswami, "Optical layer survivability: a service perspective", IEEE Communication Magazine, pp. 104 –113, March 2000.

[10] Ch. Li, R. Ramaswami, "Automatic fault detection, isolation, and recovery in transparent all-optical networks", Journal of Lightwave Technology, Vol. 15, No. 10, pp.1784 –1793, October 1997.

[11] T.M. Chen, T.H. Oh, "Reliable services on MPLS", IEEE Communication Magazine, pp. 58 – 62, December 1999.

[12] K. Owens, S. Makam, V. Sharma, B. Mack-Krane, "A path protection/restoration mechanism for MPLS networks", IETF draft, draft-chang-mpls-path-protection-02.txt, July 2000.

[13] V. Sharma, B.M. Crane, S. Makam, et al., "Framework for MPLS-based recovery", IETF draft, draft-ietf-mpls-recovery-frmwrk-01.txt, November 2000.

[14] L. Mo, "General considerations for bandwidth reservation in protection", IETF draft, draft-mo-mpls-protection-00.txt, work in progress, July 2000.

[15] D. Haskin, R. Krishman, "A method for setting an alternative label switched path to handle fast reroute", IETF draft, draft-haskin-mpls-fast-reroute-05.txt, November 2000.

[16] P. Demeester, M. Gryseels, A. Autenrieth, et.al., "Resilience in Multilayer Networks", IEEE Communication Magazine, August 1999.

ICMP Covert Channel Resiliency

Kristian Stokes, Bo Yuan, Daryl Johnson, and Peter Lutz
Department of Networking, Security, and Systems Administration
B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York 14623

***Abstract*—The ICMP protocol has been widely used and accepted as a covert channel. While the ICMP protocol is very simple to use, modern security approaches such as firewalls, deep-packet inspection and intrusion detection systems threaten the use of ICMP for a reliable means for a covert channel. This study explores the modern usefulness of ICMP with typical security measures in place. Existing ICMP covert channel solutions are examined for compliance with standard RFCs and resiliency with modern security approaches.**

I. Introduction

The Internet Control Message Protocol (ICMP) [1] is designed to provide feedback about problems in the communication environment. ICMP relies on the basic support of IP as part of a higher level protocol. Due to this dependency, both ICMPv4 and ICMPv6 exist for both versions of IP. Many message codes exist within ICMP to properly diagnose network problems and traffic flow. ICMP messages are sent under many different circumstances such as an unreachable destination or general congestion control on the network. Simple network troubleshooting utilities such as ping and traceroute utilize explicit ICMP messages to gather information about a network.

Covert channels often refer to a hidden information stream and more specifically, hidden streams embedded in IEEE 802 networks. Lampson [2] originally defined covert channels under a number of categories such as storage, timing, termination, resource exhaustion and power. Most covert channels that involve the use of ICMP are largely storage channels where unused fields are utilized for covert communication. ICMP as a covert channel provides many benefits due to the overall simplicity. Only several fields exist within most ICMP messages which enables quick implementation and simple channel setup/teardown. The idea of using ICMP as a covert channel is to use an atypical communication protocol rather than ordinary TCP or UDP. This will have a smaller footprint across the network and may go unnoticed by network administrators and traffic analyzers. What really makes the ICMP protocol a viable covert channel is the use of data fields or payloads within certain messages. By generating packets based on specific message codes and embedding the actual covert channel message in the data field enables ICMP to serve as an alternate use for covert channels. These simple factors enable ICMP to be considered as stealth traffic.

II. Current Countermeasures

Countermeasures exist within the ICMP covert channel realm, however they come at a cost. As with all aspects of security, tradeoffs exist.

Blocking all ICMP traffic from entering the network prevents all aspects of ICMP communication, including covert channels. This methodology may not be acceptable due to the loss of network troubleshooting abilities. Blocking ICMP communications at a central firewall through the use of ACL mechanisms may not solve the problem. The capability to send ICMP messages in an intranet environment may still exist.

Blocking specific ICMP messages from entering the network. This methodology also incorporates the use of segmenting incoming and outgoing connections. If this countermeasure is enabled, both parties involved in the covert channel can develop fuzzing techniques to modify ICMP messages used for communication or party initiation.

Restricting the size of ICMP packets. By blocking large ICMP packets, an ICMP message with an extensive data field will be dropped and perceived as a crafted packet with a malicious or unknown payload. Large ICMP packets can be used to test a network for proper handling of large packets. An adversary can also overcome this limitation by fragmenting their ICMP covert channel to smaller packets.

Traffic normalization. Traffic normalization techniques such as state preservation will reply and generate new packets on the senders behalf. The normalizer, typically a firewall, will serve as a proxy, rebuild messages and construct new payloads. This activity essentially disrupts ICMP covert channel communication by stripping out the message payload. State preservation requires significant computational power and may not scale to particular environments with high traffic loads.

III. Related Work

Research in this area focuses on existing solutions currently available for ICMP covert channel communication. Loki, an ICMP tunneling back door application, tunnels remote shell commands in ICMP echo reply / requests and DNS query / reply traffic. This proof of concept was originally published in Phrack World News [3] to demonstrate the vulnerabilities within these protocols. This implementation is very easy to deploy and thusly carries a security risk of ICMP tunneling.

Another implementation named Ping Tunnel [4], is focused on reliably tunneling TCP connections using ICMP echo request and reply packets. This tool can be with an outside

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_87, © Springer Science+Business Media B.V. 2010

proxy to tunnel traffic using ICMP messages back to the requesting client. The exterior proxy serves all TCP requests and forwards the data back to the client via ICMP echo reply. Although this solution can be viewed as subverting typical network communication, it is also important to realize other potential uses for ICMP covert communication.

ICMP-Chat [6] implements a basic console-based chat mechanism that utilizes ICMP packets for communication. A unique aspect of this solution incorporates the use of an encrypted data field using AES-256. By implementing an encrypted payload this further secures the covert channel, but adds increased suspicions on an abnormal ICMP payload.

With the increasing awareness of IPv6, covert channel tools further expanded into the realm of ICMPv6. A tool v00d00N3t [7] was developed to operate specifically over IPv6. This tool focuses on the infancy of IPv6 protection technology. Dual IPv4/IPv6 routers are utilized to route IPv6 traffic to send messages and files using ICMPv6. Useful information is embedded into fields other than the data field. For instance, the sequence number tells the covert channel receiver if it should read the packet and ID field is used to identify how many bytes out of the payload to read.

Additional defensive research has been performed to further limit the capabilities of ICMP covert channels. A Linux kernel module was developed to scan ICMP messages for specific signatures such as passwd, root, etc, ls and dir [5]. If these signatures were detected, the ICMP message was scrubbed by zeroing out the data field while being processed by the network stack. This technique is similar to normalizing traffic, however the action of data scrubbing is performed on the end nodes. This solution can also be deemed as a very computational intensive process involving bi-directional deep-packet inspection. The added network processing overhead may not be acceptable for performance reasons.

With these current solutions outlined, little information is available for actual survivability of the tools and general ICMP covert channel message resiliency with common security appliances. The use of intrusion detection, intrusion prevention and firewalls are commonplace within a production environment. To fully understand the capabilities of ICMP as a covert channel, it must contain a moderate amount of resiliency in modern networks.

IV. Existing Solutions

ICMP-Chat [6] is based on a simple console-based chat interface which uses ICMP packets for communication. Features of this solution include the use of AES-256 encryption for encrypting chat data across the channel. ICMP-Chat provides several protection mechanisms such as password protecting the channel using SHA-256 and supporting the ability to change usage of ICMP codes within the application. This allows for mobility between ICMP codes to further obfuscate the channel.

Ping Tunnel [4] allows a user to tunnel TCP connections to a remote host using ICMP echo request and reply packets. This is achieved through the use of a proxy serving as the remote host. This solution can be utilized as a covert channel when the operating network is heavily restricted.

Each solution will be tested and further explored in the following sections in terms of resiliency with modern security devices such as firewalls and intrusion detection systems.

V. Experiment

A minimalistic test environment was created in a lab environment using VMware Workstation. Two CentOS workstations are used as the endpoints of the covert channel. A Checkpoint NGX R65 firewall separates the network between internal and external clients. The internal network simulates a small business with a firewall and IDS in place. The external network simulates an Internet node communicating to the business. ICMP communication is permitted to the interior node, all remaining traffic is dropped at the firewall.

The Checkpoint firewall serves as a network layer inspection engine to detect fields and patterns in the network layer. This firewall provides sufficient inspection for a network layer covert channel.

A Snort IDS [8] is deployed inline on the network listening in promiscuous mode. All traffic passing through the internal network is captured and analyzed by the IDS. The standard Snort ruleset was updated to the latest version from Sourcefire.

The basis for this experiment focuses on the resiliency of ICMP as a viable covert channel. ICMP-Chat and Ping Tunnel will be used in this environment to test resiliency against a modern firewall and IDS. Resiliency will be examined for each tool on connection establishment, connection integrity and overall covert design.

VI. Results

Given the provided experiment, both ICMP-Chat and Ping Tunnel are examined for ICMP resiliency. Each tool was installed and tested for functionality on both internal and external workstations.

A. *ICMP-Chat*

By default, ICMP-Chat uses echo reply packets for primary communication. Figure 1 illustrates the network topology of the experiment. When comparing standard operating system echo reply packets to echo reply packets generated by the ICMP-Chat application many parameters follow standard RFC compliance.

The standard CentOS echo reply, shown in Figure 2 consists of several key focus areas such as packet size and data content. A total of 98 bytes are captured on the wire with a 56 byte data field. The data field also standardizes on the following sequence of characters: "!#$%&'()*+,-./01234567". For a perfect covert channel, the ICMP communication should match this similar format to reduce detection.

An ICMP-Chat echo reply session was initiated, sent with the message "test" and captured in Figure 3.

Major differences with the reply packet in Figure 3 centers on the overall packet size of 318 bytes with a data size of 276 bytes. The dramatic increase in data size contributes to

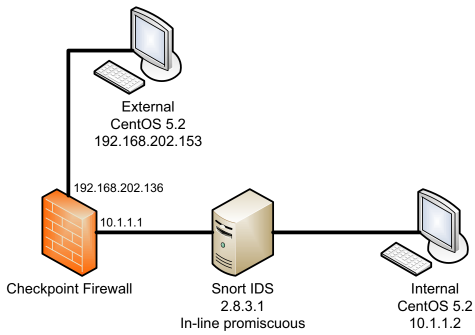


Fig. 1. ICMP-Chat covert channel environment

```
      Frame Frame 2397 (98 bytes on wire, 98 bytes captured)
      Ethernet II, Src: Vmware_3a:02:dd (00:50:56:3a:02:dd), Dst:
Vmware_3a:07:f8 (00:50:56:3a:07:f8)
      Internet Protocol, Src: 10.1.1.2 (10.1.1.2), Dst: 192.168.202.153
(192.168.202.153)
      Internet Control Message Protocol
  Type: 0 (Echo (ping) reply)
  Code: 0 ()
  Checksum: 0x2568 [correct]
  Identifier: 0x2448
  Sequence number: 1 (0x0001)
  Data (56 bytes)
      Data: CB9E0B4AEC62080008090A0B0C0D0E0F1011121314151617

      0000  00 50 56 3a 07 f8 00 50 56 3a 02 dd 08 00 45 00   .PV:PV:.E.
      0010  00 54 84 94 00 00 40 01 5f d0 0a 01 01 02 c0 a8   .T.@._.......
      0020  ca 99 00 00 25 68 24 48 00 01 cb 9e 0b 4a ec 62   .%h$H..J.b
      0030  08 00 08 09 0a 0b 0c 0d 0e 0f 10 11 12 13 14 15   .............
      0040  16 17 18 19 1a 1b 1c 1d 1e 1f 20 21 22 23 24 25   ....... !#$%
      0050  26 27 28 29 2a 2b 2c 2d 2e 2f 30 31 32 33 34 35   &()*+,-./012345
      0060  36 37                                             67
```

Fig. 2. Standard CentOS echo reply

```
      Frame 491 (318 bytes on wire, 318 bytes captured)
      Ethernet II, Src: Vmware_3a:02:db (00:50:56:3a:02:db), Dst:
Vmware_3a:07:f7 (00:50:56:3a:07:f7)
      Internet Protocol, Src: 192.168.202.153 (192.168.202.153), Dst: 10.1.1.2
(10.1.1.2)
      Internet Control Message Protocol
  Type: 0 (Echo (ping) reply)
  Code: 0 ()
  Checksum: 0x8857 [correct]
  Identifier: 0x0000
  Sequence number: 1280 (0x0500)
  Data (276 bytes)
      Data: 00000000000000000000000000000000000000000449569F8...

      0000  00 50 56 3a 07 f7 00 50 56 3a 02 db 08 00 45 00   .PV:...PV:....E.
      0010  01 30 eb f9 00 00 40 01 f7 8e c0 a8 ca 99 0a 01   .0....@.........
      0020  01 02 00 00 88 57 00 00 05 00 00 00 00 00 00 00   .....W..........
      0030  00 00 00 00 00 00 00 00 00 00 00 00 00 00 44 95   ..............D.
      0040  69 f8 98 69 2a cf 87 6c 41 eb bf 95 14 0c 7e b8   i..i*..lA.....~.
      0050  af 0c d4 17 ad d0 2e 76 b8 75 cf 29 a4 8f 4c f7   .......v.u.)..L.
      0060  75 fe 22 86 a4 96 a6 14 0d 03 3c 2a 0d c5 95 9d   u.".......<*....
      0070  ca cc 56 1d 39 f0 21 ab 1d a9 76 0c 32 ec 71 3a   ..V.9.!...v.2.q:
      0080  11 66 17 46 0a dc 5d ce 74 41 f8 bf e7 ee a1 1e   .f.F..].tA......
      0090  c2 b9 88 07 6a 18 f2 88 3c 3d 5d 29 a4 b9 49 ac   ....j...<=])..I.
      00a0  bd c3 91 ec 82 da 2b d9 eb 4d b7 8e 54 44 ef 04   ......+..M..TD..
      00b0  d4 91 4d 30 d7 91 c8 36 0f 92 d1 47 4b 50 e5 1f   ..M0...6...GKP..
      00c0  8e 68 c6 90 1d 9c a0 46 bb b0 2f 64 17 ed 60 ca   .h.....F../d..`.
      00d0  d9 81 89 b7 30 08 9d 80 c6 f6 cd 4e 7c 0f 82 67   ....0......N|..g
      00e0  15 04 fe e6 d9 8f 7b 20 a6 12 be 2c 90 0e b5 2a   ......{ ...,...*
      00f0  8d 16 56 60 d0 85 f8 60 1f 3d 54 25 d4 71 3b 95   ..V`...`.=T%.q;.
      0100  8e 24 82 68 b4 64 0d be 00 35 92 67 6e 4d 46 e0   .$.h.d...5.gnMF.
      0110  40 85 f2 a7 89 76 23 75 d4 72 7a 7e 94 4e 09 3e   @....v#u.rz~.N.>
      0120  5f 29 43 c7 ce ec 45 a1 f6 1a 76 7c 79 af 87 dc   _)C...E...v|y...
      0130  d1 63 26 fd 28 48 b3 63 ca 43 5f 82 94 00         .c&.(H.c.C_...
```

Fig. 3. ICMP-Chat echo reply

an abnormal attribute for typical ICMP traffic. Large packet size decreases the covertness of this solution. If the network restricts large ICMP packets, ICMP-Chat will likely be blocked.

Firewall Resiliency. Session initiation of ICMP-Chat with the Checkpoint firewall must follow request and reply structure. By default, ICMP-Chat uses echo reply packets. Similar to stateful firewall inspection, the Checkpoint firewall expects an echo request then permits an echo reply. If both internal and external nodes use the echo replies, the firewall does not permit the communication. If the communication is changed to an echo request and reply structure the Checkpoint firewall permits the communication. This similar structure must be followed for continued communication; an echo request must be received before an echo reply is permitted to traverse the firewall.

IDS Resiliency. The Snort IDS was unable to detect abnormal ICMP traffic when conducting the covert channel. This further confirms that abnormal ICMP packet sizes are not added to the standard IDS ruleset. Given that the data field is encrypted, this adds to the level of complexity needed for IDS detection in covert channels.

B. *Ping Tunnel*

Ping Tunnel [4] serves as a covert tunneling tool to disguise TCP traffic in ICMP request and reply packets. The basis of this technique is to disguise traffic as a wrapper protocol and bypass specific TCP filtering. This is achieved through the use of an external proxy to translate transmitted ICMP client packets back to standard TCP packets. Unlike ICMP-Chat, Ping Tunnel strictly uses ICMP echo request and reply packets for communication. Connections are very similar to TCP in that lost packets are resent as necessary to allow for reliability. Multiple connections are permitted through the use of the ICMP identifier field. The identifier field is included within the standard Ping Tunnel packet format and should not be confused with the ICMP sequence number field.

The example test environment, illustrated in Figure 4, was designed to allow the client node to establish an SSH tunnel to an Internet based server. A known proxy address was provided to the client and listens on a local port which tunnels all traffic via ICMP echo request/reply packets. Once the ICMP proxy was established, an SSH connection was initiated from the client to the localhost port.

Upon establishment of the connection, the standard SSH handshake can be viewed in captured ICMP echo request/reply packets.

The packet capture in Figure 5 shows a standard OpenSSH version handshake embedded in the data field of a ICMP echo reply.

Firewall Resiliency. The Checkpoint firewall permitted Ping Tunnel traffic largely due to the adherence to the ICMP echo request and reply structure. If this request and reply format is continued throughout the communication, the traffic will go largely unnoticed. Unsolicited connections will be dropped by the firewall. Similar to ICMP-Chat, the Checkpoint firewall did not specifically block large ICMP packets.

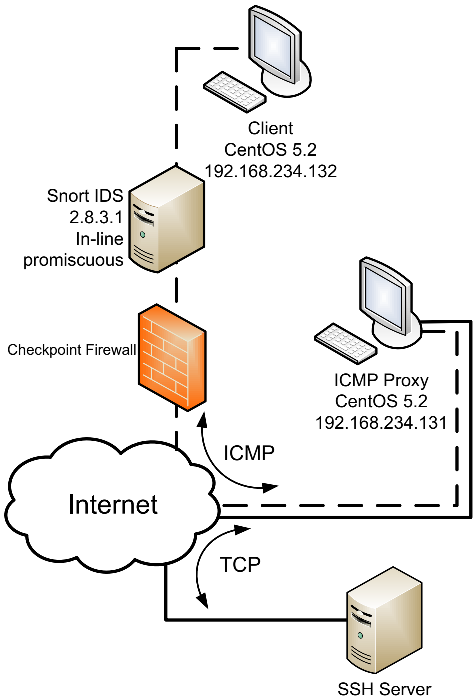


Fig. 4. Ping Tunnel covert channel environment

```
No.     Time         Source                 Destination       Protocol
10  0.012010    192.168.234.131       192.168.234.132       ICMP

Frame 10 (92 bytes on wire, 92 bytes captured)
Ethernet II, Src: Vmware_3a:02:dd (00:50:56:3a:02:dd), Dst: Vmware_3a:02:db
(00:50:56:3a:02:db)
Internet Protocol, Src: 192.168.234.131 (192.168.234.131), Dst: 192.168.234.132
(192.168.234.132)
Internet Control Message Protocol
    Type: 0 (Echo (ping) reply)
    Code: 0 ()
    Checksum: 0x195d [correct]
    Identifier: 0xe510
    Sequence number: 0 (0x0000)
    Data (50 bytes)

0000  d5 20 08 80 00 00 00 00 00 00 00 00 80 00 00 02   . ..............
0010  00 00 00 00 00 00 00 15 00 00 e5 10 53 53 48 2d   ............SSH-
0020  32 2e 30 2d 4f 70 65 6e 53 53 48 5f 35 2e 31 0d   2.0-OpenSSH_5.1.
0030  0a 20                                             .
        Data: D520088000000000000000000800000020000000000000000015...
```

Fig. 5. Ping Tunnel OpenSSH handshake

IDS Resiliency. The inline Snort IDS also failed to detect this covert channel. Again, abnormally sized ICMP messages are overlooked given this situation. It is possible to tailor a custom ruleset to detect this activity, but the default ruleset fails to recognize frequent, abnormally large ICMP messages.

VII. CONCLUSION

Based on the experimental findings of this study it is very clear that a simple ICMP based covert channel can easily subvert many modern security appliances if general ICMP traffic is permitted. Administrators and security researchers should be aware of the capabilities of a seemingly helpful protocol. Current covert channel tools are widely and freely available for use across a number of platforms.

Blocking all ICMP traffic may not be an acceptable business practice in many cases. Steps can be taken to further reduce the risk of an ICMP covert channel. Limiting the overall packet size of ICMP messages may disrupt communications. Ensuring unsolicited ICMP messages are dropped at the perimeter can assist in preventing channel establishment, but as investigated, this can be easily circumvented by following a request and reply structure.

An IDS can be further improved by monitoring ICMP traffic flow. Ping Tunnel generated abnormally sized ICMP packets and often produced constant or bursting traffic to the proxy node. IDS technology can be implemented to flag abnormal traffic for further investigation. The traffic analysis can be compared to constant streams of ICMP traffic resulting in varying packet sizes to a single destination.

Even with many security mechanisms and precautions in place, covert channels may still exist. This activity is essentially what a covert channel fundamentally strives to be, an undetectable channel of communication.

REFERENCES

[1] Postel, J. "INTERNET CONTROL MESSAGE PROTOCOL." RFC 792. Internet Draft Submission Tool. 30 Apr. 2009.

[2] Lampson, B. W. 1973. "A note on the confinement problem." Commun. ACM 16, 10 (Oct. 1973), 613-615.

[3] Daemon9. 1997. "LOKI2." Phrack Magazine, Vol. 7 (51) http://www.phrack.com/issues.html?issue=51&id=1.

[4] Stodle, Daniel. "Ping Tunnel." http://www.cs.uit.no/ daniels/PingTunnel.

[5] Singh, Abhishek, et al. 2003. "Malicious ICMP Tunneling: Defense against the Vulnerability." In: Lecture Notes in Computer Science, Vol. 2727, Springer Berlin / Heidelberg, pp. 226-236.

[6] Muench, Martin J. 2003. "ICMP-Chat." http://icmpchat.sourceforge.net/index.html.

[7] Murphy, R. P. 2006. "IPv6 / ICMPv6 Covert Channels." DEFCON 14. https://forum.defcon.org/archive/index.php/t-7588.html.

[8] Sourcefire. "Snort IDS." http://www.sourcefire.com.

Chip-to-chip Free-space Optical Interconnection Using Liquid-crystal-over-silicon Spatial light modulator

Sadia Sultana, Ferdous Mohammed Shahriar, Mohammad Khalad Hasan
Department of Computer Science & Engineering
Shahjalal University of Science and Technology
Sylhet-3114, Bangladesh.
sadiasust@yahoo.com, ferdousshahriar@yahoo.com, khalad-cse@sust.edu

Abstract—**Free-space optical chip-to-chip interconnect offers a promising solution to bandwidth requirement. To move a light beam from a chip to another chip we need beam steerers which will steer the beam to desired direction. Different beam steering technologies have already been demonstrated to control the direction of light beam. These include different types of prisms, micro-electro-mechanical systems (MEMS), and Opto-VLSI processors. But all of those technologies have the limitations of having high optical losses & low signal-to-noise ratio. But the spatial light modulator (SLM) has several advantages over these. The large number of independently controllable elements in SLMs is expected to offer low optical loss as they are dynamic, controllable, and repeatable without moving parts. In this paper, we have proposed the concept of chip-to-chip optical interconnect that employs vertical cavity surface emitting lasers (VCSEL) and photo detectors (PD) in conjunction with two SLMs and Hologram design. This model for chip-to-chip free space optical interconnects uses SLM as beam steerer which is easy to control. Using our proposed model, we can have low optical losses, low crosstalk. This proposed chip-to-chip optical interconnection model can be applied in processor-memory and processor-processor communication.**

Keywords-Diffraction, Electrode, Hologram Design, Liquid Crystal, Spatial light modulator.

I. INTRODUCTION

In electrical interconnect, density is a great problem. The smaller elements on a chip operate faster. As the number of elements in a chip increases, this requires more connections, and placing all of these connections on a 2-D surface is a difficult optimization problem. Over short distances, for high frequency clock distribution or high-speed data transfer optical interconnect will be the most feasible solution. Reconfigurable free-space optical interconnect modules employing polarization-selective diffractive optical elements in combination with a liquid crystal based polarization controller have previously been reported by Goulet, et al. However, these modules result in high optical losses especially when the number of output ports increases [1]. In next, using Opto-VLSI processors as beam steerers and multicasters for reconfigurable interchip optical interconnection was reported [1]. Optical beam direction compensating system for board-to-board free space optical interconnection in high-capacity ATM switch was proposed [2]. Free-space switch demonstrator using reconfigurable hologram design was reported [3]. Free-space board-to-board optical interconnect structure employing a single ferroelectric liquid-crystal spatial light modulator in conjunction with a free space optical polarizing beam splitter, halfwave plates and collimating lenses has been reported [4]. Chip-to-chip optical interconnection using MEMS mirrors was experimented by Tod Laurvick [5]. But still a complete & feasible technology for chip-to-chip free-space optical interconnection does not exist. In our model we have proposed the use of SLMs as beam steerer in conjunction with Hologram design. As both of these have several advantages over other technologies, we expect to have better performance using this model.

The rest of the paper is organized as follow: at first we introduce the functions of SLM & Hologram design in section II and III respectively. In Section IV, our proposed model for chip-to-chip interconnection is presented. Section V concludes the paper.

II. THE LIQUID CRYSTAL OVER SILICON SPATIAL LIGHT MODULATOR

A spatial light modulator (SLM) is a device that imposes some form of spatially-varying modulation on a beam of light. Usually, an SLM modulates the intensity or phase of the light beam, or both the intensity and the phase simultaneously. In our model the SLM only changes the phase of light without changing the intensity and relation of polarization state. A silicon backplane phase modulating SLM acts as a beam-steerer. It consists of a liquid-crystal-over-silicon (LCOS) backplane, a layer of nematic liquid crystal, and a cover glass. The LCOS backplane is an array of aluminum pixels which serve as both reflective mirrors and electrodes. Each electrode is an independently controllable pixel. The modulation is done according to the alignment of liquid crystal which is controlled pixel by pixel. This silicon backplane also contains addressing circuitry for the pixels. The liquid crystal layer is considered to be divided into different block each corresponds to an electrode pixel. With each electrode an electric field can be applied from the LCOS backplane across the liquid crystal block to achieve optical modulation. The cover glass is coated with a transparent electrode. As light enters the LC

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_88,

layer through cover glass, an electric field is applied to the layer to change the refractive index that depends on molecular orientation of liquid crystal layer block. The index change induces a phase-only modulation of the input light. It is then reflected from the aluminum mirror and passes back through the liquid crystal layer [7]. Multilevel phase modulation can be achieved using a nematic material with the correct birefringence and molecular orientation [8]. The SLM chip is mounted in a ceramic Pin Grid Array (PGA) package. This PGA package connects to a signal controlling board for receiving and sending data, power and control signals [9]. The signal is collimated by the lens onto the SLM. By controlling the deflection angle by SLM, the beam may be made to return to a selected point by collimating again by a lens.

Following figures shows a cross-sectional view of a liquid-crystal-on-silicon spatial light modulator illustrates how the liquid-crystal layer causes a voltage-dependent phase shift upon entering light. Each pixel is fully programmable, offering a high-resolution pure phase-modulating array [6].

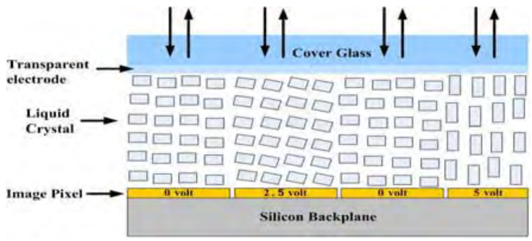


Fig. 1. A cross-sectional view of a liquid-crystal-on-silicon spatial light modulator.

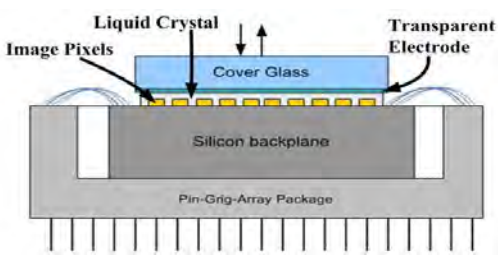


Fig. 2. Voltage-dependent phase shift upon entering light.

III. HOLOGRAM DESIGN

To route a beam to a single point binary phase hologram is used. The hologram refers to the phase modulation 2D pattern and period applied by the SLM. Usually the technique of holography is used to optically store, retrieve and process information. By using a reconfigurable hologram we can steer the beam to any desired positions in the detector plane. One of the most fundamental mechanisms of optical propagation is through the process of diffraction. In phase only hologram or grating diffraction is done without blocking photons. By using a hologram or grating to steer the light by diffraction, we can correct for any motion or misalignments and create an adaptive optical interconnect. The SLM must maintain DC balance during the display of each corrective pattern. [8]. A controller algorithm is used to design a suitable adaptive binary holograms to calculate the amount of phase modulation required for alignment of the light beam. In order to achieve active alignment we need to generate holograms in real time thus we can use a non-iterative algorithm [8]. When the light beam incidents on the SLM then the phase modulation causes it to split into a number of reflected beams and a phase modulation pattern is created. Each beam corresponds to a different diffraction order of the phase modulation of which only one is created intentionally to carry the signal into the selected output point. The other diffraction orders are unwanted and they may lead to crosstalk and back reflection. The first-order diffraction efficiency of a binary-phase hologram with LC layer thickness, and molecular tilt angle θ is given by equation (1), where Δn is the birefringence and λ is the operating wavelength [3].

$$\eta = \sin^2 2\theta \sin^2\left(\frac{2\pi\Delta nd}{\lambda}\right) \qquad (1)$$

The output angle of each diffraction order, measured from the optical axis is given by equation (2), where Λ is the period of the phase modulation, usually known as the hologram period, and m is an integer that identifies a particular diffraction order and λ is wavelength of the laser light that is used [3].

$$\sin\phi_{out} = \sin\phi_{IN} + \frac{m\lambda}{\Lambda} \qquad (2)$$

Fig. 3: Incident and Reflected Light Beam.

Figure 3 shows the mechanism of driving a light beam to a selected pixel by using phase hologram.

Fig. 4 Multiple Beam steering using Diffraction.

If a multicasting phase hologram is synthesized, multiple beams can be generated (Figure 4), whose intensities can be controlled by reconfiguring the phase hologram [1]. And one of our main concerns is to create multiple beams to communicate with multiple chips for parallel communication.

IV. PROPOSED MODEL

The descriptions and functions of SLM and hologram design were stated above. The complete mechanism of our proposed model is give here:

Each chip has a VCSEL & a Photo Detector (PD) with it. To communicate with another chip, each chip send light beam by using VCSEL & in the designated chip this beam of light is detected by using a PD. For each VCSEL an individual pixel of SLM is assigned. Each PD also has an individual pixel assigned to it. Light from a VCSEL is collimated and focused by using a collimating lens and it incidents on the first SLM. The addressing & other information are sent to the signal controlling board. Then the hologram pattern is generated using Fourier series analysis and the required diffraction angle is calculated by using an appropriate controller algorithm. Light beam is then phase shifted and directed toward the pixel of the second SLM which is assigned to the destined PD. The second SLM again do the same things and direct the light beam to the specific PD. A collimating lens is used to collimate and focus the light beam to the PD for a chip. The different scenarios of this model that can be implemented are as follows:

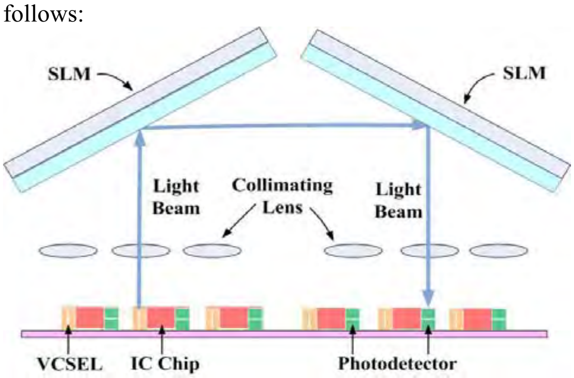


Fig. 5. 1-to-1 interconnect scenario.

Figure 5 shows the first interconnect scenario, where a VCSEL in one side connects to a PD of another side. Both SLMs are loaded with proper steering holograms to steer the beams of VCSEL #2 to PD #2.

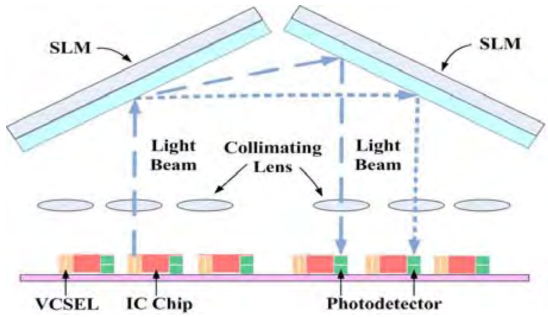


Fig. 6. 1-to-N interconnect scenario.

Figure 6 shows the second interconnect scenario, where multicasting interconnects can be established by using multicasting phase hologram. Both SLMs are loaded with proper steering holograms to steer the beams of VCSEL #2 to PDs #1 and #2.

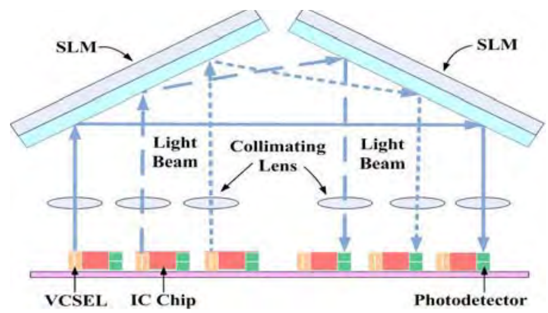


Fig. 7. N-to-N parallel interconnect scenario.

Figure 7 shows the third interconnect scenario, where parallel optical interconnects can be established simultaneously by two SLMs using the VCSELs and the PDs. Both SLMs are loaded with proper steering holograms to steer the beams of VCSEL #1, #2, and #3 to PD elements #3, #1, and #2.

V. CONCLUSION

In this paper, we have proposed a model based on the performance analysis of others concept for chip-to-chip free space optical interconnects. This uses Spatial Light Modulators as beam steerer, Hologram design, VCSELs as light sources & PD as light detectors. It has different scenarios including 1-to-1, 1-to-N & N-to-N interconnections. SLMs have the advantages of being dynamic, controllable, and repeatable without moving parts [6]. Thus it offers a good long-term reliability. The deflection angle is a function of only the hologram pattern displayed on the SLM giving repeatable and predictable performance, and as holograms are relatively insensitive to pixel errors there are excellent prospects for high yield fabrication of the devices. It is also possible to correct for optical aberrations in the system. Finally, the prospect of liquid crystal on silicon (LCOS) devices means that, in principle, the necessary control electronics could be integrated on to the device to make a very compact system [4].

There are two key challenges of implementing this model. First, as the SLM is pixellated, so the possible hologram periods must be an integer multiple of the pixel pitches. Second, the unwanted diffraction orders that can lead to crosstalk and back reflection must be routed and controlled appropriately, otherwise these will greatly reduce the performance of the system. Finally, SLM consists of a layer of nematic liquid-crystal, but the nematic devices have a significantly slower switching speed than ferroelectric devices.

In this preliminary study we have considered only chip-to-chip interconnection. This proposed model of optical interconnects can also be applied in board-to-board communication.

REFERENCES

[1] Muhsen Aljada and Kamal E. Alameh , Yong-Tak Lee and Il-Sug Chung, "High-speed (2.5 Gbps) reconfigurable inter-chip optical interconnects using opto-VLSI processors", Centre for MicroPhotonicSystems, Electron Science Research Institute Joondalup, WA, 6027, Australia; pp 3-8, 2006 Optical Society of America.

[2] Katsuhiko Hirabayashi, Tsuyoshi Yamamoto, Shigeki Hino, Yoshitaka Kohama, and Kouta Tateno, "Optical Beam Direction Compensating System for Board-to-Board Free Space Optical Interconnection in High-Capacity ATM Switch". Journal of Lightwave Technology, Vol. 15, No. 5, May 1997.

[3] W. A. Crossland, I. G. Manolis, M. M. Redmond, K. L. Tan, T. D. Wilkinson, M. J. Holmes, T. . Parker, H. H. Chu, J. Croucher, V. A. Handerek, S. T. Warr, B. Robertson, I. G. Bonas, R. Franklin, C. Stace, H. J. White, R. A. Woolley, and G. Henshall, "Holographic Optical Switching: The "ROSES" Demonstrator"; Journal of Lightwave Technology, Vol. 18, No. 12, pp 2,3, December 2000.

[4] C.J. Henderson, B. Robertson, D. Gil Leyva*a*, T.D. Wilkinson, D.C. O'Brien, G. Faulkner Cambridge Univeristy Engineering Department, Trumpington St, Cambridge, CB2 1PZ, UK Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, UK. "Control of a free space adaptive optical interconnect using a VCSEL array transmitter and FLC SLM for beam steering"; cjdh2@cam.ac.uk, pp 1.

[5] Tod Laurvick**,** "Chip-To-Chip Optical Interconnection Using MEMS Mirrors", March 2009.

[6] http://www.laserfocusworld.com/articles/255505.

[7] http://jp.hamamatsu.com/en/rd/publication/scientific_american/common/pdf/scientific_0806.pdf.

[8] T. D. Wilkinson, C. D. Henderson, D. Gil Leyva and W. A. Crossland, "Phase modulation with the next generation of liquid crystal over silicon technology", Journal of Materials Chemistry, pp 1-4, June 2006.

[9] Boulder Nonlinear Systems**, "**32-channel Spatial Light Modulator Driver Subsystem**",** Boulder Nonlinear Systems, pp 2, June 1, 2001.

GA Based Neural Network Retraining Using Adaptive Genetic Operations

C.I.Bauer; H.Yu; B.Boffey
Faculty of Computing, Engineering and Technology, Staffordshire University, Stafford, England
C.I.Bauer@staffs.ac.uk; H.Yu@staffs.ac.uk; boffey@liverpool.ac.uk

***Abstract*—Within cellular systems prediction has proven to be a potential solution to enhancing the handover procedure to guarantee constant Quality of Service to mobile users. By using historical route information, the future movement of mobile devices is predicted in advance with the aim to reserve resources prior to arrival of the device in a new cell. However, as the traffic patterns of devices in this environment change over time this needs to be taken into consideration when designing a prediction system. Some mechanism has to be provided to address this issue. This paper presents a Genetic Algorithm (GA) based retraining scheme using adaptive layer-based genetic operations for a Neural Network (NN) based movement prediction systems in a cellular environment to enhance system performance in the presence of changing traffic patterns. Experimentation has shown that using an adaptive layer-based GA approach can provide a significant improvement to the predictive properties of the NN based prediction system in the presence of changing movement patterns.**

***Index Terms*—Cellular systems, genetic algorithms, handover, prediction, neural network, retraining**

I. Introduction

Genetic algorithms (GAs) have over the years been applied to a variety of problems with great success. One area of application is the use of GAs for initial training and tailoring the weights of neural networks (NN) to a given problem. Research in this field has clearly shown that the use of GAs can be beneficial, providing equally good results in the fraction of the time required by conventional NN training techniques through backpropagation algorithm. Hence, they hold the potential to widen the field of applications NNs can be applied to. Traditional applications of NNs are based on systems that are likely to encounter information with the same characteristics that are fed to the NN for processing. During training of a NN this information is used to adjust the weights accordingly to produce the required response. While in most NN based systems the characteristics of the information are not likely to change much, retraining is rarely a consideration. However, depending on the problem to which a NN is applied, this may not necessarily be the case. This paper presents a GA based retraining scheme using layer-based adaptive mutation and crossover operations for a NN based movement prediction systems in a cellular environment to enhance system performance under changing traffic patterns. Starting with a review of related research work combining GAs and NNs, the following sections will provide details of the experimental results of a GA with layer-based adaptive mutation and crossover operations for NN retraining to the given problem.

II. Related Work

GAs have been investigated and applied to a large variety of problems over the years due to their ability to search the solution space efficiently delivering optimal solutions to problems to be solved. This ability has been exploited by numerous authors through combining GAs with other approaches including NNs and applications found in publications are manifold. The concept of NN retraining within this is not entirely new and was also recognised previously by authors in other subject areas, including NN-based prediction systems resulting in a variety of schemes suggested for different scenarios. Schemes that fall into this category include the approach by Nastac and Matei [1], as referred to and detailed in [2] and [3], that proposes a NN retraining scheme based on scaling the reference weights by a defined scaling factor. They can then be used as initial weights for a new training cycle to improve the network accuracy and the adaptable NN model for recursive non-linear traffic predictions of communication networks proposed in [4] that is based on estimating the NN weights to change them according to the current network conditions. However, the majority of authors concentrate on using GAs for implementing feature selection algorithms to reduce the amount of training data fed to the NN to shorten training times and enhance the network response as proposed in [5]; for the initial training of the NN weights as discussed in [6] and [7]; to identify a suitable network design for a given problem and periodically redesign the network structure and determine the network weights such as the load forecasting model [8] and the short-term traffic flow forecasting scheme in [9] for intelligent transportation systems (ITS). Although there are a large number of GA-based NN schemes for a variety of problems, the possibility of using GAs to retrain a network with a fixed network structure to maintain performance levels in the presence of changing conditions appears so far to have been mostly overlooked.

III. Genetic Algorithm Retraining Scheme

The GA based NN retraining scheme discussed in this chapter is a continuation of the work discussed in [10] and assumes that the NN based movement prediction systems introduced is the same publication is used to predict the future movement trajectories of mobile devices. For this prediction system the network is trained with mobile device route information that has been accumulated over time and reflects the devices regular movement patterns. Using this for training enables the NN-based movement prediction

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_89, © Springer Science+Business Media B.V. 2010

system to predict future movement, through the information in the distribution probability matrix – maintained for each cell within the network grid – and the previous movement step, with a certain level of accuracy. However, as the movement patterns of mobile devices change over time this will affect the prediction success rate of the system. Without retraining with new route information that reflects the changed movement pattern of mobile devices the prediction success rate will drop over time. As full retraining using traditional training mechanisms such as the backpropagation algorithm can be very time-consuming and inconvenient at times, GAs can offer an alternative solution to the problem, or at least provide a temporary measure to allow the system to continue with acceptable performance levels until full retraining is possible. This algorithm could be employed once the prediction success rate has dropped below a defined threshold to improve the prediction success rate.

When applying GAs to a problem there are two main requirements as defined in [11]: firstly, a suitable way to represent the problem/solution needs to be found. Secondly, in order to be able to evaluate the solutions of each generation, a fitness function needs to be defined that allows for selecting the best solutions from each generation with which to continue the evolutionary process. In addition to this parameters such as population size, number of generations, how to implement the genetic operations and apply them to the defined chromosomes as well as any additional parameters that arise due to the scenario-specific implementation have to be identified. In the case of the given scenario using NNs in combination with GAs, two parameters are indirectly given through the design of the NN and the way its performance is evaluated. As the aim of the system is to successfully predict the next visited cell of a mobile device, the prediction success rate is used as the fitness function to evaluate individuals of a generation. In addition to the fitness function, elitism was used to carry across the best solution to the next generation without alteration. To represent the problem, the weight information provided through the trained network was encoded in a chromosome as depicted in the diagram below.

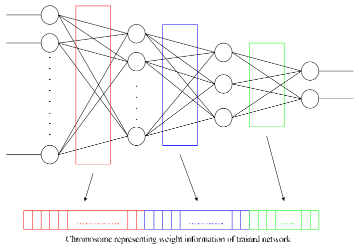


Fig. 1. Conversion of NN weight information into chromosome

This representation was chosen because the vital information of a trained network is stored in the weights for the connections between the different nodes and these are the parameters that determine the prediction success rate for a given traffic pattern. When the traffic patterns change this is the information that needs updating to maintain prediction performance levels.

To create the initial generation to start the GA the weight information of the trained and out-of-date NN was used as a basis, as it can be argued that the ideal set of weights for the new traffic patterns are very close to the new optimal solution. Hence this holds the potential to further reduce training times. To generate the individuals of the initial population, each gene of the original set of weights was modified by a small, randomly selected value.

The remaining parameters used for the GA were a mutation rate of 1/N (with N being the length of the chromosome), single point crossover with a crossover rate of 50% using the roulette wheel selection algorithm with elitism as well as the population size of 180 individuals. The number of individuals selected for the next generations was set to 50% of the population size. These parameters were established through initial simulations to optimise the GA retraining approach. The simulations were run with 10, 20, 50 and 100 generations and the results for the parameter set optimised for the retraining approach compared against the standard settings as defined by DeJong [12]. The changing traffic patterns were simulated through data sets with varying levels of known regular route information ranging between 100%–50% in steps of 10% the NN had encountered during training. The remaining route information within the data sets consisted of unknown regular route information that had not been experienced previously. The results of the simulations can be seen in the diagram below which is showing the achieved fitness for the different number of generations averaged across the investigated data sets for ease of representation:

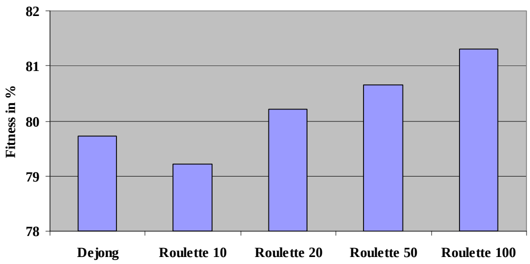


Fig. 2. Performance comparison

As can be seen from the diagram the results received for the customised parameter set using 20, 50 and 100 generations for varying levels of new unknown regular route information within the data set clearly outperformed the standard DeJong settings for Neural Network retraining. However, although the simulation results showed that the GA retrained network can improve the fitness value compared to the original network without retraining, the results obtained were lower than expected. Therefore, the characteristics of the provided route information were investigated further to identify a reason for this. The weight information of networks trained with route information of 100%, 95% and 90% of regular known routes was investigated. It was observed that with each progressing layer of hidden nodes, the weight difference of the network trained with 100% regular routes compared to 95% and 90% as well as the network trained with 95% regular routes compared to 90%, increased, as depicted in the diagram below:

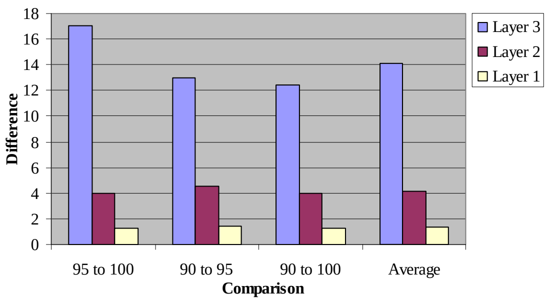


Fig. 3. Comparison of NN weight information

This suggests that when performing the genetic operations on the chromosomes it may be beneficial to distinguish between the different layers of hidden nodes contained within them. For this reason the application of the mutation and crossover operations were reconsidered to take into account the established pattern to investigate the effect of this on the GA results.

The mutation operation was hence adjusted to reflect this by varying the amount the value of a gene is altered when mutation is applied. In the experiment – if mutation was applied to a gene of the chromosome – a random value ranging between 0 and 1 was multiplied by a multiplier value and then randomly added or subtracted after being scaled based on the hidden layer the gene belonged to. To investigate different multiplier values and identify the most suitable one this value was ranged between 1 and 200. The investigation showed that a value of 30 for the multiplier produced the best results.

The crossover operation was also altered to take into consideration that within a chromosome the weight information (genes) is arranged per network layer and per node within each network layer. Hence the crossover operator was applied differently. Only genes within a randomly selected node of a particular layer after a randomly selected point within that node were crossed over as depicted below. To find the optimal crossover rate to perform the adaptive crossover operation were conducted, which identified a rate of 20 as the best value.

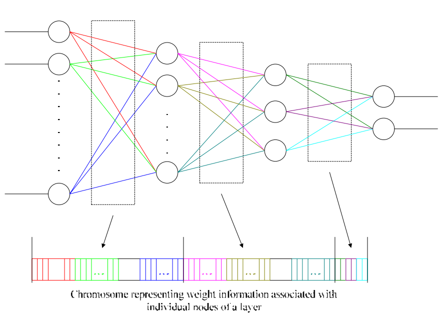


Fig. 4. Chromosome with weights encoded per network node

As in the case of the non-adaptive GA investigation, the parameters for the adaptive implementation of GA for the purpose of neural network retraining, were tested overall to allow comparison and performance analysis. With the exception of the parameters for the adaptive elements identified through the preliminary investigation, all remaining parameters remained the same as in the case of the non-adaptive GA approach. For performance comparison the adaptive GA implementation was run against the standard DeJong settings as well as the non-adaptive implementation investigated in the previous section. In order to get a good response and representative results for the comparison run with the standard DeJong settings different intermediate steps were experimented with using 10, 20, 50 and 100 generations as previously. The simulations runs were repeated three times for both the non-adaptive GA parameters and the DeJong parameters for the different training data sets ranging from 50% – 100% in steps of 10%. The results were then compared against the adaptive approach using an adaptive mutation and crossover operator as part of the scheme. The results of the simulation runs for the different data sets and training data sets are depicted below showing the performance in the same format as previously presented:

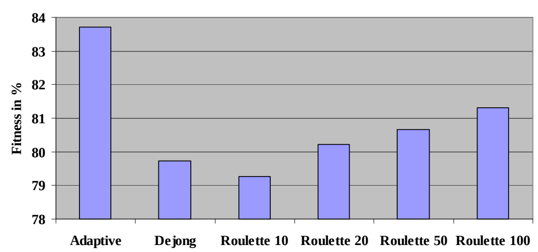


Fig. 5. Performance comparison with adaptive operations

As can be seen from the graph, the layer-based adaptive mutation and crossover have provided significant improvements to the achieved fitness. This indicates that there is a clear benefit to taking into consideration the structure of the weight information contained within a chromosome when applying the mutation operation. This can also clearly be seen in the diagram below which shows the degradation of the prediction success rate of a trained network in the presence of changing traffic patterns and the prediction success rate of the network after retraining using GA based NN retraining with adaptive operations.

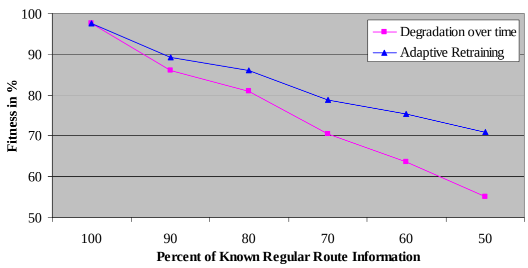


Fig. 6. Comparison adaptive GA retraining versus natural degradation

As can be seen in figure 6, the adaptive GA retraining algorithm was able to significantly improve the fitness achieved, compared to the fitness values of the simulation run without retraining. As expected the improvement is more significant for data sets with a higher percentage of new unknown regular route information as it is for data sets with a lower percentage of new unknown regular route information. A maximum of over 15% for the data set with 50% known route information was established. This is confirmation that using an adaptive GA retraining scheme is a viable solution to provide temporary relief to NN-based prediction systems in the presence of changing traffic patterns until full retraining is possible.

As the improvement in the prediction success rate is only one of the deciding factors on the success of the GA based NN retraining approach for the given problem further studies need to be conducted. To investigate whether the use of GAs for retraining would provide a time benefit for the prediction scheme, time measurements were taken and compared to the prediction success rates. As a comparison algorithm the original Backpropagation approach was used. Subject to the investigation were data sets with varying levels of known regular route information ranging from 50 – 100% in steps of 10%. To obtain representative time measurements three simulation runs were conducted and the results averaged across the three runs per data sets. Obtained simulation results are shown below:

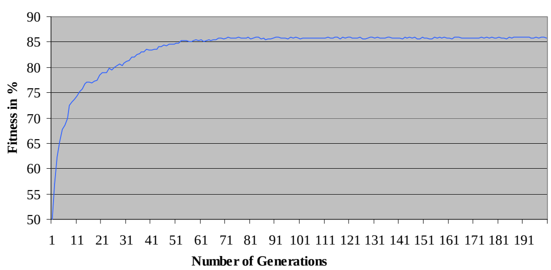


Fig. 7. GA Training Cycle

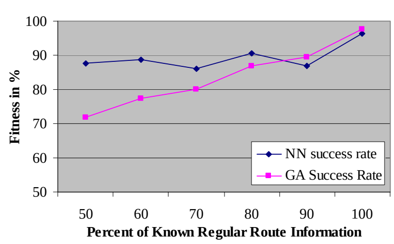


Fig. 8. Comparison NN versus GA prediction success rate

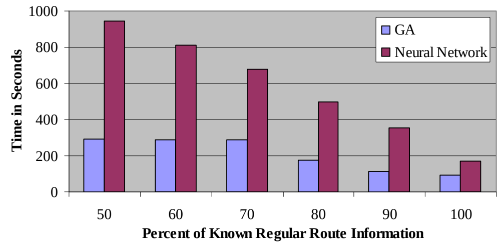


Fig. 9. Comparison Training Times Neural Network versus GA

Looking at the time measurements for the two retraining techniques in figure 9 the Backpropagation approach shows steadily increasing time requirements with a rising percentage of unknown route information within the data sets. For the GA approach there also seemed to be rising time requirements for data sets with higher levels of known route information which, after the first few data sets, however, seemed to level off. Overall seen, the retraining time requirements of the GA approach were well below those of the Backpropagation approach clearly confirming the time benefit of the GA scheme. This is down to the fact that the GA algorithm in previous experiments seemed to converge after a relatively small number of generations as can be see in figure 7 which shows the best and the average population score for an example run of the GA.

Considering the results of both graphs it can be concluded that looking at the prediction success rate alone, the Backpropagation algorithm appears to be the better choice of retraining the network. However, considering the original aim of the GA retraining, this approach has a clear advantage on the training times. As the GA retraining is only meant to be used as a temporary measure to extend the life of the Neural Network until full retraining is possible again this proves that this is a viable way to achieve this. In light of the obtained simulation results, taking into consideration the time saving and the improved prediction success rate, it can further be concluded that when using the GA approach for retraining to set a threshold value to a suitable value to benefit from the better prediction success rate achieved by the GA algorithm.

IV. CONCLUSION

In this paper a GA-based retraining algorithm using adaptive mutation and crossover to assist the NN based prediction systems in maintaining its performance levels in the presence of changing traffic patterns is presented. As simulation work has shown, the adaptive layer-based implementation of the genetic operations provided improvements to the system's capabilities to increase the prediction success rate (fitness) compared to the standard mutation genetic operations. This is a strong indication that when applying genetic operations, the structure of the weight information contained within a chromosome should be taken into consideration.

Further to the investigation of enhancing the prediction success rate through adaptive genetic operations when retraining a NN, time measurements were taken to compare the traditional Backpropagation algorithm and the GA based retraining algorithm with respect to retraining time and achieved prediction success rate. Simulation showed that looking at the prediction success rate the Backpropagation algorithm clearly outperformed the GA retraining approach. However, with respect to retraining times, the GA retraining approach showed a clear time benefit over the traditional retraining approach, although not achieving the same accuracy levels as the Backpropagation approach. As the adaptive GA based NN retraining approach is intended as a temporary measure to bridge the time until full retraining is possible, again this proves that this is a viable way to achieve this.

REFERENCES

[1] Nastac, I.; Matei, R. "Fast retraining of artificial neural networks"; Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Springer-Verlag in the series of Lecture Notes in Artificial Intelligence; 2003; pp. 458.462; ISBN: 978-3-540-14040-5.

[2] Nastac, I.; Costea, A. "A Retraining Neural Network Technique for Glass Manufacturing Data Forecasting; International Joint Conference on Neural Networks", 2004; Proceedings 2004; Volume 4; 25-29 July 2004 Page(s): 2753 – 2758.

[3] Nastac, I.; Cristea, P., "ANN Flexible Forecasting for the Adaptive Monitoring of Multi-Tube Reactor"; 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services; 14th International Workshop on Systems, Signals and Image Processing 2007; 27-30 June 2007; Page(s): 193 – 196.

[4] Doulamis, A.D.; Doulamis, N.D.; Kollias, S.D., "An Adaptable Neural-Network Model for Recursive Nonlinear Prediction and Modeling of MPEG Video Sources"; IEEE Transaction on Neural Networks; Volume 14, Issue 1; January 2003 Page(s): 150 – 166.

[5] Ichibuchi, H.; Nakashima, T.; Nii, M., "Learning of Neural Networks with GA-Based Instance Selection"; Joint 9th IFSA World Congress and 20th NAFIPS International Conference, 2001; Volume 4, 25-28 July 2001; Page(s): 2102 - 2107 vol.4.

[6] Ling, S.H.; Lam, H.K.; Leung, F.H.F.; Tam, P.K.S., "A Genetic Algorithm Based Neural-Tuned Neural Network"; IECON '03; The 29th Annual Conference of the IEEE – Industrial Electronics Society, 2003; Volume 3, 2-6 Nov. 2003 Page(s): 2423 – 2428 Vol.3.

[7] Zuo, Guoyu; Liu, Wenju; Ruan, Xiaogang, "Genetic Algorithm Based RBF Neural Network for Voice Conversion"; WCICA 2004; Fifth World Congress on Intelligent Control and Automation, 2004; Volume 5, 15-19 June 2004 Page(s): 4215 – 4218 Vol.5.

[8] Zhangang, Yang; Yanbo, Che; Cheng, K.W. Eric, "Genetic Algorithm-Based RBF Neural Network Load Forecasting Model"; IEEE Power Engineering Society General Meeting, 2007; 24-28 June 2007 Page(s): 1 – 6.

[9] Ji, Tao; Pang, Qingle; Liu, Xinyun, "Study of Traffic Flow Forecasting Based on Genetic Neural Network"; Sixth International Conference on Intelligent Systems Design and Applications – ISDA '06; Volume 1, Oct. 2006; Page(s): 960 – 965.

[10] Bauer, C.I.; Yu, H.; Boffey, B. "Adaptive Genetic Algorithm for Neural Network Retraining"; International Joint Conferences on Computer, Information, and System Sciences – CISSE '08; 5-13 December 2008

[11] Núñez, Edwin; Banks, Edwin Roger; Agarwal, Paul, "High Performance Evolutionary Computing"; HPCMP Users Group Conference 2006, June 2006; Page(s): 354 - 359.

[12] DeJong, K.A. and Spears, W.M. "An Analysis of the Interacting Roles of Population Size and Crossover in Genetic Algorithms," Proc. First Workshop Parallel Problem Solving from Nature, Springer-Verlag, Berlin, 1990. pp. 38 - 47.

Decision Support for IP Uplink and Routing Choice

A Whole View of the Internet

Peng Su
Department of Computer Science and Engineering
University of Bridgeport
221 University Avenue, CT 06604, USA
pengsu@bridgeport.edu

Zhengping Wu
Department of Computer Science and Engineering
University of Bridgeport
221 University Avenue, CT 06604, USA
zhengpiw@bridgeport.edu

***Abstract*—Although Border Gateway Protocol (BGP) is broadly used among autonomous systems (ASes), the topology of peer autonomous systems is often mysterious to some Internet Service Providers (ISPs) or administrative domains. With unknown routing policies of BGP neighbors and their uplink ASes, it is uneasy to make a proper design when building a multihoming environment. Sometimes it may even cause congestion problems. Some mistakenly announced prefixes of ASes over the Internet will continuously increase routing table and deteriorate policy-based routing. This paper analyzes the adjacency relationships of ASes over the entire Internet and their announced prefixes. Using this information, our prototype system can automatically identify the more appropriate access point for individual ASes, depict the relations among ASes, and provide decision support to manage the improper address spaces over the Internet.**

Keywords-BGP; autonomous system; routing policy; adjacency relationship, topology

I INTRODUCTION

As described in [1] and [8], the Border Gateway Protocol is an inter-autonomous-system routing protocol used between Internet Service Providers, domains, regions, nations and etc. BGP exchanges reachability messages between neighbors or peers, maintains three information databases such as Adj-RIB-In, Adj-RIB-Out, Loc-RIB, as well as provides the routes that BGP has selected using local routing policies. Furthermore, this protocol supports Classless Inter-Domain Routing (CIDR [9]) and Variable Length Subnet Mask (VLSM [10]), which makes BGP peers free to aggregate IP addresses to an appropriate scale and easy to advertise proper IP prefixes not limited by classes.

A routing policy is a set of rules that regulates the traffic as well as adjacency relationships between routers and other hosts in term of route exchanges and protocol interactions. Summarily, in a routing policy [6], a customer can define:

- A collection of prefixes the router will receive or not from others.
- A collection of prefixes the router will advertise or not to others.
- Redistribution of prefixes between various routing protocols.

These three items can be defined in incoming policies, outgoing policies, route-maps and access lists.

II PROBLEMS CAUSED BY POLICIES

Traditionally, routing policies between autonomous systems are designed following local autonomous system's interests and not opened to other External Border Gateway Protocol (EBGP) neighbors. Therefore, one BGP customer or one ISP may be blind to the area out of their own AS and neighboring ASes. After balancing interests and needs, one institution may believe that some "desired" uplinks could become its major outgoing and incoming interfaces because that might work better. Other uplinks will only be used as backup links. However, this is not always true. For example, from the view of an AS, one BGP user may know little about adjacency relationships out of the local domain, this BGP user may unwittingly suffer from packet loss and delay problems, when congestion occurred between its uplink AS and the destination AS. Figure1, 2, and 3 together explain this issue. Here, Figure1 illustrates the normal situation of packet forwarding between AS1 and C1 (C1 is a customer needs uplinks). C1 chooses the link to AS2 as its default route to AS1 and the Internet, the link to AS3 as its backup link. However, when some congestion happens between AS 1 and AS2 (as illustrated in figure 2) the packets to C1 will be partially dropped. Especially, when the traffic load between C1 and AS 2 is light, C1 will not switch its main link to AS3 automatically and thus suffers from this remote congestion. This situation will be more often and more complicated when there is more than one AS between AS2 and AS1 (as illustrated in figure 3).

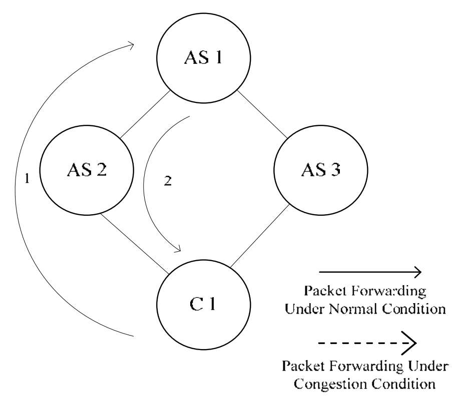


Figure1. Packet forwarding in normal condition with blindness to the topology of uplink AS

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_90, © Springer Science+Business Media B.V. 2010

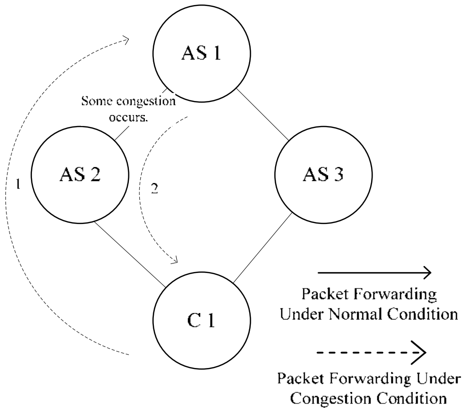


Figure 2. Packet forwarding in congestion condition with blindness to the topology of uplink AS

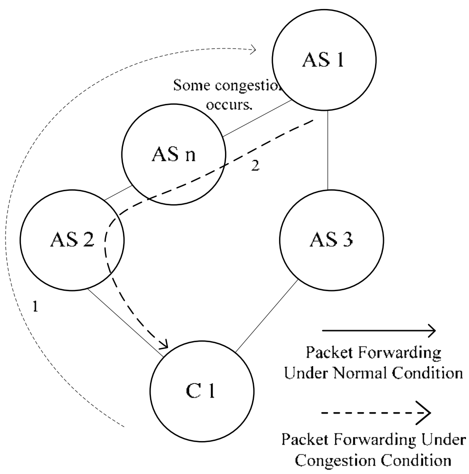


Figure 3. Packet forwarding in congestion condition (multi-hop) with blindness to the topology of uplink AS

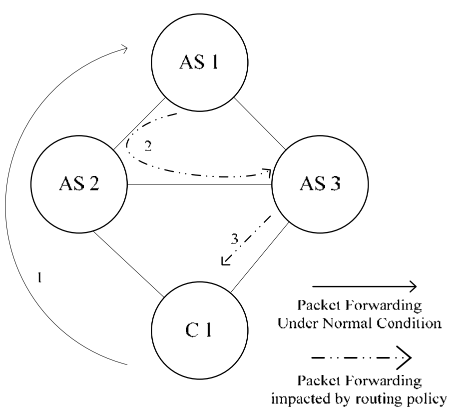


Figure 4. Packet forwarding impacted by routing policy and blindness to the topology of uplink AS

In another example, when the forwarding action is influenced by routing policies or routing control mechanisms [3], there will be unnecessary hops between one AS and its destination. Generally, it is difficult for a customer to detect and improve the quality of packet transferring in its uplink ASes, because some ISPs not only aggregated the address spaces, but also filtered the detecting packets on their equipments. As illustrated in figure 4, C1 needs to send packets from AS2 to AS1 where C1 chooses AS2 as its main uplink, and AS3 as its backup link. But AS1 is imperceptible to C1, so packets between AS1 and C1 will go through the path C1-AS2-AS1, and AS1-AS2-AS3-C1. It is clear that packets are affected by routing policies and suffer from unnecessary hops. Besides, if there exist one or more autonomous systems between AS2 and AS3, the forwarding action will become more complicated with more uncertainty. At the same time, packet loss, loop, and long delay of data will appear at C1 with this design.

However, things will change when the customer, C1, knows the topology outside its local area. The customer, C1, can use AS3 as their major uplink, whereas the connection to AS2 as a backup. Of course, many customers may not be aware of the importance of topology outside their local AS. But this topology is helpful, sometimes necessary. Moreover, nowadays many BGP routing tables have already reached 300,000 entries, which require extra hardware recourses to support effective and efficient routing. This situation may be caused by following reasons:

（1） The increasing needs of IP addresses
（2） Announced overlapping prefixes
（3） Some announced prefixes whose network masks are more than 24

Along with the requirements of network expansion, routing entries in a router's memory will grow. However, the situation (1) and (2) can be avoided when some ISPs and domains pay more attention on the whole view of the Internet. For example:

Situation (1)

route-views.oregon-ix.net> show ip bgp 123.252.128.0/18
BGP routing table entry for 123.252.128.0/18,
Paths: 2905 702 6453 4755 17762

Situation (2)

route-views.oregon-ix.net> show ip bgp 123.252.129.0/24
BGP routing table entry for 123.252.129.0/24
Paths: 2905 702 6453 4755 17762

These two examples show one overlapping prefix announced by AS 17762. Although this method may lead to traffic redundancy, the increasing number of BGP routing table entries makes convergence [7] time rise continuously.

Situation (3)

route-views.oregon-ix.net> show ip bgp 199.77.194.254
BGP routing table entry for 199.77.194.254/32
Paths: 3333 1103 20965 6509 10764 19401 10490

In this situation (3), one prefix whose network mask is 32 emerges in the BGP table. This situation may be caused by some configuration errors or specific needs. Nevertheless, it should be avoided whatever the considerations are. Therefore, it is unwise to show every autonomous system's route entries on a router and check them one by one. Analyzing BGP entries, making a statistics from all ASes, and then checking route entries as well as making policies following the statistics are more feasible and practical. The analysis result from the statistics is called the whole view of the Internet.

III ACCESS CHOICE TO THE INTERNET

Access choice to the Internet can be reached through the database recording the whole view of the Internet, which contains two tables. The table "AS" collects the neighborhood of ASes and the table "ROUTES_BY_AS" stores the prefixes announced by each AS. Table "AS" consists of three fields, AS_ID, Nei_num, and Nei_as, where AS_ID includes the AS numbers over the entire Internet, Nei_num is the number of EBGP peers, and Nei_as field describes the EBGP neighbors of the autonomous system in AS_ID. Also, in the table of "ROUTES_BY_AS," the fields of "Route_number" and "Route_entries" describe the route entry number as well as the content of each prefix. Like table "AS," this table contains one field named "AS_ID" too, which acts as the index for this table.

Each line of table "AS" contains binary relations of autonomous systems. $R=(X, Y, G(R))$ represents a relation between X and Y, in which X consists of AS number identified in AS_ID field, Y consists of the AS number(s) of the EBGP neighbors peering with the AS in AS_ID field, and G(R) means the graph of R. Equally, $R \subseteq X \times Y = \{<X, Y> \mid X \in AS\_ID,\ Y \in Nei\_as\}$. For a better way to access the Internet, we can assume $R'=(Y, Z, G(R'))$, and $R' \subseteq Y \times Z = \{<Y, Z> \mid Y \in AS\_ID,\ Z \in Nei\_as\}$. Here, Y may contain one or many neighboring AS numbers. Simultaneously, in the lines indexed by AS_ID of Y, each element of Y can also have one EBGP peer or several EBGP peers. The collected EBGP peers or peers of every element of Y is Z. If putting this action forward, the recursion will continue until the last binary relation contains a proper Internet Exchange Point (IXP) or local area IXP. Then the shortest path of ASes to some IXP may be a better choice to access the Internet. For example, $X= \{15430\}$, $Y= \{34, 981\}$, then R is the Cartesian product of X and Y. $R= \{<X, Y> \mid X \in (15430),\ Y \in (34, 981)\} = \{(15430, 34), (15430, 981)\}$. $R'_{Y=34}= \{<Y_{34}, Z_{Y=34}> \mid Y \in (34),\ Z_{Y=34} \in (2158, 7018)\} = \{(34, 2158), (34, 7018)\}$. $R'_{Y=981}= \{<Y_{981}, Z_{Y=981}> \mid Y \in (981),\ Z_{Y=981} \in (36, 1209)\} = \{(981, 36), (981, 1209)\}$. In this example, AS 7018, AT&T WorldNet Services, could be a proper IXP, therefore, AS 15430 may choose AS 34 as its main link and another one as the backup link.

Each line of table "ROUTES_BY_AS" is indexed by AS_ID, and includes the prefixes originated from this AS. Let $M_i = \{i.prefix1, i.prefix2 \ldots \mid i = AS\_ID\}$, $M' \subset M_i$, $N_j = \{j.prefix1, j.prefix2 \ldots \mid j = AS\_ID\}$, and $N' \subset N_i$. Then $M' \cap N' = \varnothing$.

IV A WHOLE VIEW OF THE INTERNET

The construction of a whole view of the Internet is discussed in this section.

A. Algorithm description

The data source, containing 38 BGP peers, reaches 9 million lines of route entries. To discover the adjacency relationships and prefix information, these two following algorithms are designed to analyze all BGP routing entries, regulate them, form topology among ASes, and put each prefix's information into database by each AS number. Here, the algorithm for analyzing AS adjacency relationships is named "Analysis Algorithm," as illustrated in figure 5. The major goals are:

(1) To filter peer relationships of ASes and remove unhelpful messages, like the metric and weight.

(2) To get rid of the repeated AS number generated by AS prepending (append AS_NUM to AS path), search the database by AS number, and update the database.

```
Input
 Whole BGP routing tables from 38 EBGP peers

Output
 Database BGP,   table "AS"

Phase 1: regulate strings

  Read routing table into memory;
    While (i<= some certain number)
       {if (!resource is at the end)
            {/* Because the amount of resource is huge,
              the certain number is one part of the resource.
              */
          truncate the route entries, and put them into
          string array;
          }
         else break;
         i++;
        }
   Return string array in the certain pattern;

Phase 2: analyze strings and update database

 Connect database
  while (! the string array is at the end)
    {
     if (! pointer of one element is at the end)
       {
         Check and remove overlapped ones;
         if (this AS number_n  exists in table "AS")
             {
               locate the line, check to see whether it
               contains this one (AS number)
               before/behind AS number_n;
                 if (exist)
                    continue;
                 else
                     {
                      update this line and continue;
                     }
                 }
         else
              {
                 insert one line, and update this line;
               }
        }
     else
        {
           read next string from the array;
         }
     }
  return the disposed string array;
Close database connection;
```

Figure 5. Analysis Algorithm for regulating and analyzing BGP routes

The second algorithm is designed to regulate and classify the BGP prefixes. It is named "Class Algorithm," as illustrated in figure 6. The tasks of this set of programs can be described as:

```
Input
  Whole BGP routing tables from 38 EBGP peers

Output
  Database BGP, table "ROUTES_BY_AS"

Phase 1: regulate strings
    Read prefixes into memory;
      while (i<= some certain number)
       {
         if (!resource is at the end)
           {
            truncate entries to the beginning of prefixes;
           }
          else
           break;
         i++;
       }
  Load lines into array and return prefix array;

Phase 2: classify route entries by AS number and update
database
  Connect database;
     while (!prefix array meets the end)
       {
         read one line from prefix array;
          if (this AS number exists in table )
            {
              update database by increasing the route
              entries number and adding the prefix in this
              line;
            }
         else
            {
              insert one new line in table "
              ROUTES_BY_AS";
              update database by adding the prefix in this
              line;
            }
       }
```

Figure 6. Class Algorithm for regulating and classifying BGP routes

B. The statistics of AS adjacency relationship

Based on the analysis result, it is not difficult to make a statistics of AS adjacency relationships and their scale. The following figures, tables, and data are used to illustrate the current status of adjacency relationships among autonomous systems, their topology, and the characteristics of the entire Internet.

Here, there are two data granularities on figure7 in that the number of ASes with more than 500 EBGP peers is much less than that with 100 to 500 EBGP peers. At the same time, these ASes whose EBGP neighbors range from 100 to 500, are much less than the number of ASes with less than 100 EBGP peers. Figure 7 illustrates the relationship between the number of ASes and their EBGP peers. X axis is the number of EBGP peers, and Y axis is the number of ASes.

As the figure 7 shows, the number of AS entities whose EBGP neighbors are more than 500 is only 11, and the number of AS entities whose EBGP peers are more than 1000 is 5. Comparing with the total number of autonomous systems, 30583, these 11 AS entities count for only 0.036%. Autonomous systems belonged to this area mainly contain backbone IXP, such as 701, 7018, 3356, and etc.

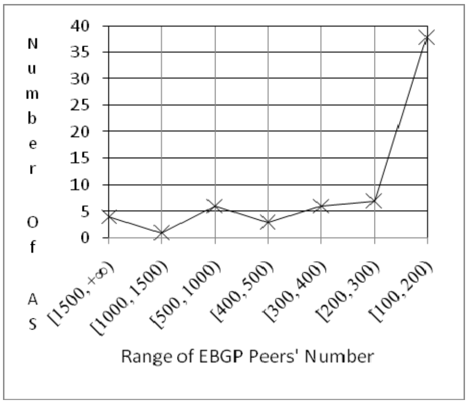


Figure 7. Distribution of Autonomous Systems whose EBGP peers are more than 100

When comparing to the total number of ASes, the numbers of autonomous systems go up as their EBGP neighbors fall. For example, stub ASes take more than half of the total AS entities. Its percentage reaches 57.27%. In the view of autonomous systems, these stub areas do not respond to data relay, while those ASes that have over 10 EBGP peers are mainly responsible for the connectivity of other areas, though they take a comparatively small part, 2.67%, over the entire Internet.

C. The statistics of prefixes provided by database

When analyzing the prefixes originated from different autonomous systems, the algorithm can also handle irregular address spaces on the Internet. The common characteristic of these address spaces is their masks are over 24. For example, the number of prefixes whose net masks are 32 reaches 20 and the number of those entries whose net masks are 25 is 110. Although these entries could be used among autonomous systems, it is unwise to receive and advertise them in routing policies. Therefore, these "bugs" are entirely and explicitly examined by the algorithm, which can provide policy decision for routing control.

V DECISION SUPPORT SYSTEM

A. Decision of access

As mentioned above, traditionally when constructing a network by BGP, the top factors in consideration [5] are scalability, stability, robustness [2], and business interests. However, the whole adjacency relationships of uplink ASes could easily be neglected, not only because some designers have not recognized the importance of neighborhood outside their uplink ASes, but also these areas are blind to network designers. This paper tries to demonstrate the significance of knowing the whole adjacency relationships of ASes over the entire Internet, including local uplink ASes. By the database of the whole view of the Internet, some designers can depict the topology of their outgoing interface, discover which ones are appropriate IXP ASes, and avoid remote congestions as well as redundant hops impacted by routing policies. Suppose the situation of figure3 has occurred and AS1 is one backbone IXP. C1 does not know the topology of other ASes either because of receiving aggregated routes or default route configurations [8]. When C1 sets its main link to AS2 and undergoes some remote congestion or oscillations [4], this AS will not automatically switch its link to AS3. However, if putting this situation to the support decision system, the fruits will be like this:

Key = AS1
Neighbor number = n
Neighbor information of AS1: <AS1, ASn>, <AS1, AS3>, <AS1, ASj> … <AS1, ASk>

Key=C1
Neighbor number = 2
Neighbor information of C1: <C1, AS3>, <C1, AS2>

Key=AS2
Neighbor number=2
Neighbor information of AS2: <AS2, C1>, <AS2, ASn>

Key=AS3
Neighbor number=2
Neighbor information of AS3: <AS3, C1>, <AS3, AS1>

Key=ASn
Neighbor number=2
Neighbor information of AS3: <ASn, AS1>, <ASn, AS2>

Let ∀A and ∀B ∈ {Neighbor information of C1, AS2, AS3, ASn}, suppose that x A ∩ B if and only if x A and x B, and let M = A ∩ B, x=AS1, then (x∈M) → (∀m ∈M, x∈m) → AS1 ∈ Neighbor information of AS3 OR AS1 ∈ Neighbor information of ASn.

From the calculation of ASes' adjacency relationships, it is obvious that the congestion may occurred either in <AS1, ASn> or in <AS2, ASn>. Furthermore, AS3 is closer to AS1 from the AS-PATH aspect. Thus, changing main link of C1 will solve the problem. For another question, as illustrated in figure 4, although C1 may not be sensitive to the impacts of AS2's routing policies, after checking the topology between source and destination, C1 still can avoid the redundant hops by switch its main link to its backup link. Therefore, by introducing the decision support system using the whole view of the Internet, these two previous issues can be resolved. Besides, in the view of IXP, there will be similar issues to AS1, as described in figure 3 and figure 4. While the difference between C1 and AS1, in this scenario, is that the situations of IXP would be more complex and more difficult to deal with. However, following the normal forms explained in the previous example, the adjacency relationships, EBGP peer numbers, route entry scales make the relationships among neighbors and the entire Internet topology clearer. Moreover, when take the ranks of ASes, the stub areas, and stability of ASes into account, this decision system is able to provide more useful ranking information to make decisions.

B. The future of stub AS

According to the discovery of whole adjacency relationships of ASes over the Internet, the percentage of stub AS over the Internet reaches 57.27%, so that the star topology is still dominant and prevalent in the Internet. But as a reality, this is an unstable structure, for below reasons:

（1） Many stub ASes use only a default route to other ASes, whereas their unique uplink ASes could arise congestion, redundant hops, as well as packet loss issue. As a result, the customer stub ASes suffer from them.

（2） Even though one stub AS uses more than two links to the uplink AS, these links functions only load balance [12] and failover, which will not be helpful to the previous issues.

（3） In many cases, the traffic between two stub ASes is huge, which hold the most bandwidth of these areas. However, these packets have to go through the hub area and affected by traffic policies and unexpected routing questions.

The first and second questions may occur when some uplink ASes (maybe some ISPs or domains) conceal the topology by summary action or stub areas only use default route. Then these stub ASes are blind to the topology outside local area. Nevertheless, in our decision support system, the adjacency relationships of stub ASes and other IXP are obvious. Thus, as described in the previous section, stub ASes can get the nearest IXP information, at least in the aspect of AS-PATH, and then make a proper decision for the network design.

Nowadays, it is pretty common for the third question over the Internet. Consequently, traffic has to endure unnecessary hops between stub areas. Figure 8 and 10 illustrate this issue. Figure 9 and 11 provide a solution.

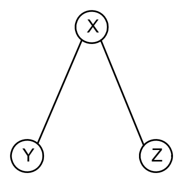


Figure 8. Unnecessary hops between two stub ASes and one hub AS

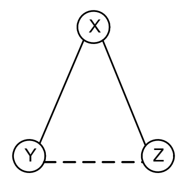


Figure 9. Refined network architecture between stub ASes

Figure 8 presents a simple "hub and spoke" topology, where X is the hub AS, and stub area Y as well as Z work as spoke. When X forwards the packets from Y to Z, the hops between Y and Z is 2. However, when the decision support system shows that AS Y and AS Z connect to the same hub AS - AS X, a more appropriate solution emerges, that is to add a link between Y and Z, as shown in figure 9.

This situation could be more complicate when introducing more hub ASes into this scenario, as figure 10 shows.

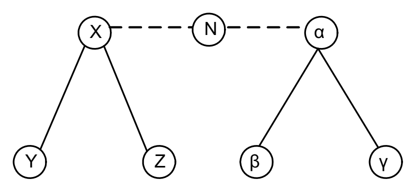


Figure 10. Unnecessary hops between two stub ASes and various hub

Let H(i, j) defines the hops between two different ASes. Suppose the traffic between AS Y and AS β holds a great amount of each AS. When Y needs to access β, then the least hops between from Y to β are:

$$\sum_{X}^{N} H(X,N) + \sum_{N}^{\alpha} H(N,\alpha) + 2$$

Therefore, each hop along the road from Y to β could influence the quality of transmission. As mentioned above, after knowing the adjacency relationships between Y and β, some solutions, as figure 11 shows, could be either connecting to the AS α to which β links or directly accessing β by adding a new neighbor to β.

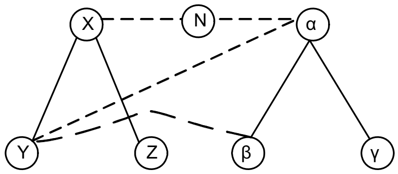


Figure 11. Refined network architecture between stub ASes and various hub ASes

Apparently, in both situation showed in figure 9 and 11, hops are greatly reduced to 1or 2, and traffic between source and destination is free from routing policies, damping issues, broadcast flood, and convergence questions.

C. Building a system free from improper address space

Since a lot of overlapping and irregular prefixes existing on the Internet, some convergence problems and security issues may increase, although some of them exist for backup functions. Especially, some domains have to receive these route prefixes for the reason that it is impossible to check every prefix redundant condition and their lengths on a device. Therefore, by checking with our decision support system, the overlapping address spaces and improper address will be on the desk. Thus a method to make routing policies and improve devices' performance is provided.

VI DISCUSSION

Fortunately, there are several BGP analysis tools available for researchers, network designers and Inter Service Providers. Most used systems among these open software are "BGPath" [13], "AS ranking" [14], and "BGP inspect" [15]. These tools have their special focus individually.

As for "BGP inspect," this service provides some main IXP's newly updated routing entries, such as ATT, Sprint, UUNET and so on. For instance, customers can submit their queries like "Global Summary Queries" according various IXP, and obtain the corresponding results. Of course, it is very useful for the network designers and researchers to track those new prefixes and making an appropriate routing policy, but when designing a network, researching the topology of ASes or ranking the ASes, it cannot help.

"AS ranking" is a supportive tool that can sort autonomous systems based on AS number, name, country, ISP information. However, this software can only provide ranking utility without a topology view of the entire Internet. Besides, it is slow, 10 minutes for producing top 50 AS rankings.

From the aspect of topology view, the "BGPath" is the popular software to show some part of the adjacencies by the customers' instructions. Comparing with "BGPath," the decision support system illustrated in this paper depicts a whole topology over the Internet and routing entries of each AS, not some part of the current network. Besides, some subject autonomous systems, such as stub areas cannot be included on "BGPath" interface. Of course, it is neat, but not enough. Hence, a complete and accurate decision system based on the relations of total autonomous systems, their routing prefixes as well as the statistics is required.

VII CONCLUSION

At the beginning of this paper, some common issues from Internet Service Providers and domains are examined. Because of the common use for prefixes aggregation, security mechanisms and default route configurations, some BGP users are blind to the entire topology outside their autonomous systems and the prefix's appropriateness on the entire Internet, which make problems above unsolvable. Therefore, there is an idea to build the whole view of the Internet, including adjacency relationships among ASes and prefixes announced from them. Then two algorithms to discover the adjacency relationships and the announced route entries are explained. After that, a decision support system based on the discovery is illustrated in the part V, which focuses on researching the adjacency relations among ASes, designing a new and more efficient network, and improving the current quality of service. Then the application of this decision support system on actual network design is illustrated, which includes designing secure policies, uplink AS choices for Internet Service Providers and stub AS optimization.

REFERENCES

[1] Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4). RFC 4271, *"The Internet Society,"* 2006

[2] K. Sriram, O. Borchert, O. Kim, and P. Gleichmann, and D. Montgomery, "A Comparative Analysis of BGP Anomaly Detection and Robustness Algorithms," *Proceedings of the Cybersecurity Applications and Technology Conference for Homeland Security (CATCH), Washington D.C.*, pp. 30-36, March 3-4, 2009

[3] Osamu Akashia, Kensuke Fukudab, Toshio Hirotsuc and Toshiharu Sugawarad, "Policy-based BGP-control architecture for inter-AS routing adjustment," *Computer Communications Volume 31, Issue 13, 15 August 2008, Pages 2996-3002*, pp. 2997-3001, 2008

[4] Alex Fabrikant, Christos H. Papadimitriou, "The complexity of game dynamics: BGP oscillations, sink equlibria, and beyond," *Symposium on Discrete Algorithms Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pp.5-7, 2008

[5] Yi Wang, Michael Schapira, Jennifer Rexford, "Neighbor-specific BGP: more flexible routing policies while improving global stability," *Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, pp. 224-227, June 2009

[6] Randy Zhang, Micah Bartell, "BGP Design and Implementation," *Cisco Press,* March 2003

[7] S. Murphy, "BGP Security Vulnerabilities Analysis," RFC 4272, January 2006

[8] D. McPherson, "Experience with the BGP-4 Protocol," RFC 4277, *Cisco Systems,* January 2006

[9] V. Fuller, "Classless Inter-Domain Routing (CIDR)," RFC1519, September 1993

[10] T. Pummill, "Variable Length Subnet Table For IPv4," RFC1878, December 1995

[11] Bruce Eckel, "Thinking in Java, 4th edition," *Prentice Hall*, February, pp 157-566, 2006

[12] X Liu, L Xiao, "Inbound Traffic Load Balancing in BGP Multi-homed Stub Networks", *The 28th International Conference on Distributed Computing Systems*, pp 369-372, 2008

[13] A. Campisano, L. Cittadini, G. Di Battista, T. Refice, C. Sasso. "Tracking Back the Root Cause of a Path Change in Interdomain Routing". *IEEE/IFIP Network Operations & Management Symposium*, 2008

[14] Xenofontas Dimitropoulos, Dmitri Krioukov, Bradley Huffaker, kc claffy, George Riley, "Inferring AS Relationships: Dead End or Lively Beginning," *Networking and Internet Architecture*, 2005

[15] D. Blazakis, M. Karir, and J. S. Baras, "BGP-Inspect - Extracting Information from Raw BGP Data," *10th IEEE/IFIP Network Operations and Management Symposium, Vancouver, Canada, April 3-7,* 2006.

Differential Cryptanalysis of Blow-CAST-Fish

Xiaoling Sun[1], Meiqin Wang[2]

1. Institute of Disaster Prevention Science and Technology, Yanjiao, Hebei, 065200

2. Key Laboratory of Cryptologic Technology and Information Security,
Ministry of Education, Shandong University, Jinan, 250100

Abstract. **Blow-CAST-Fish is a sixteen-round Feistel cipher in which the F function is a part of the private key. In this paper, we show that the disclosure of F allows perform a differential cryptanalysis against a number of rounds. We firstly identify the properties of F function with three active S-boxes to construct the 6-round differential characteristic with the probability 2^{-61} and the differential weak keys occur with the probability 2^{-12}. Then we identify the properties of F function with only one active S-box to construct the 14-round differential characteristic with the probability 2^{-49} and the differential weak keys occur with the probability 2^{-52}. With the above differential characteristics, we can attack 8-round and 16-round Blow-CAST-Fish respectively.**

1 Introduction

Blow-CAST-Fish is a block cipher designed by Krishnamurthy G.N et.al in 2008[1]. It uses the features of Blowfish [2] and CAST-128 [3].

S.Vaudenay presented the differential cryptanalysis for Blowfish based on the differential characteristics for round function with non-zero inputxor and a zero outputxor [4]. J. Nakahara Jr and M. Rasmussen presented the first concrete linear cryptanalysis on reduced round CAST-128 and CAST-256 in [6], for CAST-128, they can recover the subkey for the 4^{th} round with 2^{37} known plaintexts and $2^{72.5}$ times of 4-round CAST-128 encryption. M. Q. Wang and X. Y. Wang et.al presented the linear cryptanalytic results for 6-round CAST-128 with $2^{53.96}$ known plaintexts and $2^{88.51}$ times of 6-round encryption in [8]. Haruki Seki and Toshinobu Kaneko found the 2-round differential characteristics for F_2 round function of CAST-128 in [5], in which a non-zero inputxor resulted in a zero outputxor, then they could attack 5-round CAST-128.

As we know, there is no attack on Blow-CAST-Fish until now. Due to the S-boxes of Blow-CAST-Fish are related with the key, we will analyze the properties of F function under weak key assumptions. We will try to identify differential characteristics with high probability for round function F in which a non-zero inputxor resulted in a zero outputxor under some weak keys, then the differential cryptanalysis for 8-round and 16-round of Blow-CAST-Fish will be given. As the attack assumption in [4], we also assume the opponent knows the part of the key which describes the F function, that is the four S-boxes. And the main idea of the 8-round attack is from reference [9].

The paper is organized as follows. Section 2 introduces the description of Blow-CAST-Fish. We present how to attack 8-round and 16-round Blow-Cast-Fish in Section 3 and Section 4 respectively. We conclude this paper in Section 5.

2 Description of Blow-Cast-Fish

As a Feistel block cipher, Blow-Cast-Fish uses a block size of 64 bits, and the key size can vary from 32 bits to 448 bits. The cipher consists of two parts: a key expansion part and a data encryption part.

Key expansion converts a key of at most 448 bits into several subkey arrays totaling 4168 bytes following a scheduling scheme which works as a pseudo random generator. As this scheme is very complicated, one has to store definitely the expanded key. Our cryptanalysis is not related to the key schedule, so we don't present it in detail.

The expanded key consists of:

- 18 32-bit round subkeys $P_1, \cdots\cdots P_{18}$;
- Four arrays of 256 32-bit values which describe four S-boxes S_1, S_2, S_3 and S_4 with 8-bit inputs and 32-bit outputs.

The four S-boxes define 32-bit to 32-bit round dependent function F as follows:

(Divide the input of F into four 8-bit quarters: a, b, c and d)

Rounds：1，4，7，10，13，16

Corresponding author (email:mqwang@sdu.edu.cn)

Supported by National Basic Research Program of China (GrantNo.2007CB807902) and National Natural Science Foundation of China (GrantNo.60525201 and 60910118).

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_91, © Springer Science+Business Media B.V. 2010

$F_1 = ((S_1[a] \oplus S_2[b]) - S_3[c]) + S_4[d]$

Rounds：2，5，8，11，14

$F_2 = ((S_1[a] - S_2[b]) + S_3[c]) \oplus S_4[d]$

Rounds：3，6，9，12，15

$F_3 = ((S_1[a] + S_2[b]) \oplus S_3[c]) - S_4[d]$

Where “ ” is the bit-wise xor, “+” is the addition modulo 2^{32} and “—” is the subtraction modulo 2^{32}.

The encryption process is defined as follows:

- Split the plaintext into left and right 32-bit halves XL and XR .
- For $1 \le i \le 16$

  $XL = XL \oplus P_i$

  $XL = XL <<< Kr_i$

  (Kr_i=last 5-bit of P_i)

  $XR = F(XL) \oplus XR$

Swap XL and XR (Undo the last swap)

- $XR = XR \oplus P_{17}$

  $XL = XL \oplus P_{18}$

Recombine XL and XR as the ciphertext.

The encryption process is shown in figure 1.

3 Differential attack on 8-round Blow-Cast-Fish

For convenience, we denote the input of i^{th} round as (L_{i-1}, R_{i-1}). We consider the second round at first, the round function in round 2 is as follows:

$I = I_a \| I_b \| I_c \| I_d = (L_1 \oplus P_2) <<< kr_2$

$F_2 = ((S_1[I_a] - S_2[I_b]) + S_3[I_c]) \oplus S_4[I_d]$

The mapping $(I_a, I_b) \rightarrow S_1[I_a] - S_2[I_b]$ is a 16 to 32 bits function, so it may have a collision $S_1[I_a] - S_2[I_b] = S_1[I_a'] - S_2[I_b']$ with high probability.

Let $\delta = I_a + I_a'$, $\mu = I_b + I_b'$, assuming there is only one collision for S_1-S_2, so the probability of characteristic $\delta\mu 00 \rightarrow 0000$ from I to $((S_1[I_a] - S_2[I_b]) + S_3[I_c]) \oplus S_4[I_d]$ is 2^{-15}. So 2-round differential characteristic for Blow-Cast-Fish with the probability 2^{-15} can be produced,

R1: $(0000 \| (\delta\mu 00 >>> kr_2)) \rightarrow ((\delta\mu 00 >>> kr_2 \| 0000)$

R2: $((\delta\mu 00 >>> kr_2)\ \ 0000) \rightarrow (0000 \| (\delta\mu 00 >>> kr_2))$

We search for 2^{10} keys randomly, each key produces four S-boxes according to the key schedule of the cipher, then test the number of keys which make S_1-S_2 have a collision. We tested for twenty times to know that about half of every 2^{10} random keys will produce S-boxes which have a collision with S_1-S_2. The value of $\delta \| \mu$ is different for differential keys, and for most time there is only one value of $\delta \| \mu$ for the same weak key. So the weak keys that cause differential characteristic $\delta\mu 00 \rightarrow 0000$ for the round function in the second round occurred with probability 2^{-1}.

We will try to extend the above 2-round differential characteristic to 6 rounds. First we only consider kr_2 as zero, the other cases will be considered later.

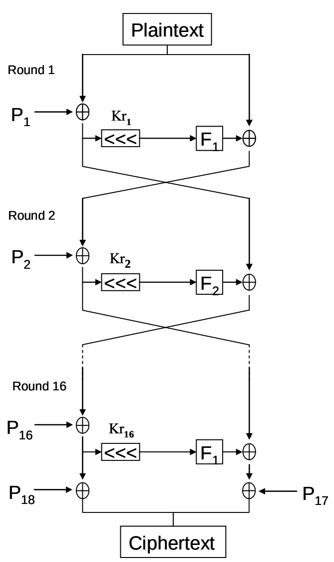


Fig 1：Feistel network showing encryption

3.1 Differential Characteristic for 6-round Blow-Cast-Fish

First, we extend the above 2-round differential characteristic to 4 rounds. The inputxor for the 3^{rd} round is zero, so the outputxor of the 3^{rd} round must be zero and the inputxor for round function F_1 in the 4^{th} round must be $\delta\mu 00$(kr_2=0), and the outputxor has not been decided. The differential characteristic of F_1 in the 4^{th} round can be denoted as follows,

$I = I_a \| I_b \| I_c \| I_d = (L_3 \oplus P_4) <<< kr_4$

$I' = I_a' \| I_b' \| I_c' \| I_d' = ((L_3 \oplus \delta\mu 00) \oplus P_4) <<< kr_4$

$\beta_1 = (((S_1[I_a] \oplus S_2[I_b]) - S_3[I_c]) + S_4[I_d]) \oplus (((S_1[I_a'] \oplus S_2[I_b']) - S_3[I_c']) + S_4[I_d'])$

In order to further extend the differential characteristic to more rounds, we only consider the case that β_1 equals to zero.

As $I \oplus I^{'}=\delta\mu00<<<kr_4$ and kr_4 varies from 0 to 31, we find out that there is always a byte of $I \oplus I^{'}$ be zero, i.e. one of the S-boxes always be non-active. So the probability of differential characteristic $\delta\mu00<<<kr_4 \rightarrow 0000$ from I to $((S_1[I_a] \oplus S_2[I_b])-S_3[I_c])+S_4[I_d]$ is 2^{-23} and the probability of 4-round differential characteristic is 2^{-38} .

We will test the proportion of weak keys together with the 6th round later.

Then we extend the above 4-round differential characteristic to 6 rounds. The inputxor of the 5th round is zero, so the outputxor of the 5th round must be zero and the inputxor of round function F_3 in the 6th round must be $\delta\mu00(kr_2=0)$, and the outputxor has not been decided. The differential characteristic of F_3 in the 6th round can be denoted as follows,

$I=I_a \parallel I_b \parallel I_c \parallel I_d=(L_5 \oplus P_6)<<<kr_6$

$I^{'}=I_a^{'} \parallel I_b^{'} \parallel I_c^{'} \parallel I_d^{'}=((L_5 \oplus \delta\mu00) \oplus P_6)<<<kr_6$

$\beta_2=(((S_1[I_a]+S_2[I_b]) \oplus S_3[I_c])-S_4[I_d]) \oplus (((S_1[I_a^{'}]+S_2[I_b^{'}]) \oplus S_3[I_c^{'}])-S_4[I_d^{'}])$

We consider the case that β_2 equals to zero.

As $I \oplus I^{'}=\delta\mu00<<<kr_6$ and kr_6 varies from 0 to 31, we find out that there is always a byte of $I \oplus I^{'}$ be zero, i.e. one of the S-boxes always be non-active. So the probability of differential characteristic $\delta\mu00<<<kr_6 \rightarrow 0000$ from I to $((S_1[I_a]+S_2[I_b]) \oplus S_3[I_c])-S_4[I_d]$ is 2^{-23} and the probability of the 6-round differential characteristic is 2^{-61} (see figure 2).

Due to S-boxes are key dependent, in order to obtain the proportion of weak keys for the 6-round differential characteristic, we need to identify the weak keys which cause the following three conditions :

1. There is a collision for S_1-S_2 with differential characteristic $\delta \parallel \mu \rightarrow 0 \parallel 0$.
2. $\beta_1=0$ for the 4th round differential characteristic.
3. $\beta_2=0$ for the 6th round differential characteristic.

We test the proportion of weak keys for the 6-round differential characteristic as follows:

1. Select 2^7 keys of Blow-Cast-Fish randomly, and identify the keys produce four S-boxes which have a collision $\delta \parallel \mu \rightarrow 0 \parallel 0$ with S_1-S_2.
2. For every four S-boxes which have a collision with S_1-S_2 and the corresponding inputxor $\delta \parallel \mu$, we exhausted the 2^{32} values of L_3 and the 2^5 values of kr_4 in the 4th round. If $\beta_1=0$, the key used is the weak key that results the differential characteristic $\delta\mu00<<<kr_4 \rightarrow 0000$ for the 4th round.
3. For the keys which satisfy both condition 1 and 2, we exhausted the 2^{32} values of L_5 and the 2^5 values of kr_6 in the 6th round. If $\beta_2=0$, the key used is weak key that results the differential characteristic $\delta\mu00<<<kr_6 \rightarrow 0000$ for the 6th round.

The key which satisfies the three conditions is the weak key for the 6-round differential characteristic.

4. Test for twenty times to get the steady proportion of weak keys.

We tested that the weak keys occurred with the probability 2^{-2}. There is at least one solution for kr_4 which satisfies the differential characteristic in the 4th round for the same keys, so the satisfied kr_4 occurred with probability 2^{-5}. Similarly, the satisfied kr_6 occurred with probability 2^{-5}.

The probability for the above 6-round differential characteristic is 2^{-61} (see figure 2), and the 6-round differential weak keys occur with the probability $2^{-5} \times 2^{-5} \times 2^{-5} \times 2^{-2}=2^{-17}$.

In order to further increase the proportion of the differential weak keys for 6-round differential characteristic, we try to test the differential characteristics with the inputxor $\delta\mu00>>>kr_2$ and zero outputxor for round functions in the 4th and 6th round as kr_2 is not zero.

As $kr_2 \neq 0$, the differential characteristics for round functions in the 4th and 6th round are $(\delta\mu00>>>kr_2)<<<kr_4 \rightarrow 0000$ and $(\delta\mu00>>>kr_2)<<<kr_6 \rightarrow 0000$ respectively. Obviously, for each value of $kr_2 \in 2^5$, there must be solutions of kr_4 and kr_6 that satisfy the differential characteristics. It means that for each non-zero kr_2, the proportion of weak keys in the 4th and 6th round is same.

So the differential weak keys which produce the 6-round differential characteristic occur with the probability $2^5 \times 2^{-17}=2^{-12}$.

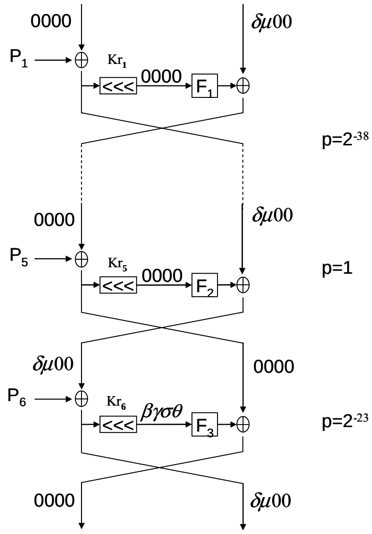


Fig2: Differential characteristic for 6-round Blow-CAST-Fish (kr_2= 0)

3.2 Differential Attack on 8-round Blow-Cast- Fish

With the differential characteristic of 6-round, we can attack 8-round Blow-CAST-Fish (see figure 3). The attack procedure is implemented as follows,

1. Choose 2^{63} values of the plaintext.
2. Loop for candidate i of kr_2(i = 0, ⋯⋯, 31):
   (a) Collect all those plaintext pairs whose XOR is δμ00>>>i at the right halves R_0. There are 2^{62} such pairs, and the number of right pairs is $2^{62}\times2^{-61}=2$.
   (b) Loop for 2^{37} values of P_{10} and kr_8 in round 8.

i. Loop for remaining pairs:

For all 2^{31} remaining ciphertext pairs

$$CL\oplus CL'=\delta\mu00>>>i$$

calculate the outputxor of round function F_2 in the 8^{th} round. If the values of P_{10} and kr_8 make the next equation holds, increase the corresponding key counter [P_{10} ‖ kr_8] by 1.

$F_2((CL\oplus P_{10})<<<kr_8)\oplus F_2((CL'\oplus P_{10})<<<kr_8)$ =CR⊕ CR'=xyzt

where C=(CL ‖ CR) and C'=(CL' ‖ CR') are a pair of ciphertexts. If the key counter is equal to 2, the subkey values of P_{10}, kr_8, and kr_2 are the right subkey with high probability.

3.3 Complexity Analysis

In this attack, k = 37 is the total subkey bits of the 8^{th} round, $p=2^{-61}$, $\beta=2^{-32}$ because the ciphertext pairs can be filtered according to their left halves XOR value δμ00>>>kr_2. For a random ciphertext pair, we denote C=(CL ‖ CR), since we have

$$F_2((CL\oplus P_{10})<<<kr_8)\oplus F_2((CL'\oplus P_{10})<<<kr_8)=xyzt$$

We can exhaustively try all the 2^{37} possible values for (P_{10}, kr_8) until the above equation holds. Given the ciphertext pair, on the average there are 2^5 solutions for the equation, so the average count of keys per analyzed pair α is equal to 2^5. Consequently,

$$S/N=\frac{2^{-61}\times2^{37}}{2^5\times2^{-32}}=2^3$$

The data complexity is 2^{63} plaintexts. The time complexity is about $2^5\times2^{37}\times2^{32}=2^{74}$ times of one-round encryption which is equivalent to 2^{71} times of 8-round encryption. The attack is efficient for Blow-CAST-Fish with key sizes more than 71 bits.

4 Differential attack on 16-round Blow-Cast- Fish

In order to increase the number of round that can be attacked, we need to improve the probability of the differential characteristic for each round. So we consider the F function with only one active S-box, see S_1 for analysis.

The mapping a→S_1[a] may have a collision $S_1[a]=S_1[a']$ under some keys. Let δ=a⊕a', kr_2=α(α＝0，1，⋯，31)，assuming there is only one collision for S_1 with input differential δ, the probability of characteristic δ000→0000 is 2^{-7}. So 2-round differential characteristic for Blow-Cast-Fish with the probability 2^{-7} can be produced,

R1:(0000 ‖ (δ000>>>α))→((δ000>>>α ‖ 0000)

R2:((δ000>>>α)   0000)→(0000 ‖ (δ000>>>α))

4.1 Differential Characteristic for 14-round Blow-Cast-Fish

We iterated the above 2-round characteristic seven times as shown on figure 4 (assuming $kr_2=kr_4=kr_6=kr_8=kr_{10}=kr_{12}=kr_{14}$＝α, βγδθ、xyzt represent undetermined values). The resulting characteristic has probability 2^{-49}.

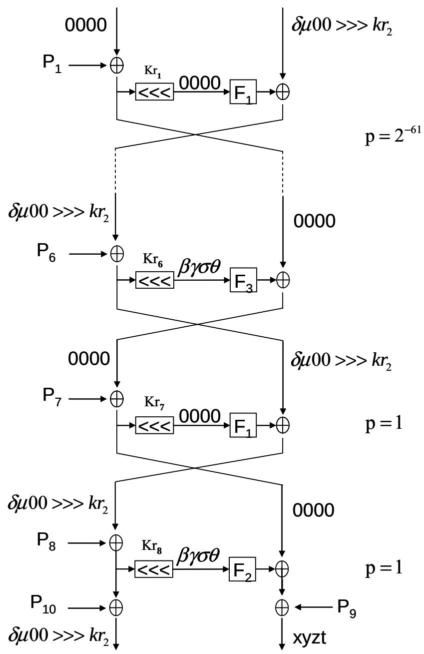


Fig 3: 8-round Blow-CAST-Fish ($kr_2 \neq 0$)

We search for 2^{20} keys randomly, each key produces four S-boxes according to the key schedule of the cipher, then test the number of keys which make S_1 have a collision. We tested for twenty times to know that at least 2^3 keys will produce S-boxes which have a collision with S_1. The value of δ is different for differential keys, and for most time there is only one value of δ for the same key. Then the weak keys which can produce the differential characteristic δ000→0000 for F functions occurred with probability 2^{-17}, $kr_2=kr_4=kr_6=kr_8=kr_{10}=kr_{12}=kr_{14}=\alpha$ occurred with probability $2^{-5\times7}=2^{-35}$, so the proportion of weak keys for 14-round differential characteristic is $2^{-35}\times2^{-17}=2^{-52}$.

4.2 Differential Attack on 16-round Blow-Cast-Fish

With the fourteen rounds differential characteristic, we can attack 16-round Blow-CAST-Fish (see figure 5). The attack procedure is implemented as follows,

1. Choose all 2^{32} values for the right halves of the plaintext R_0, which constitute a structure. Choose all 2^{20} values for the left halves of the plaintext L_0, so 2^{20} structures are produced.

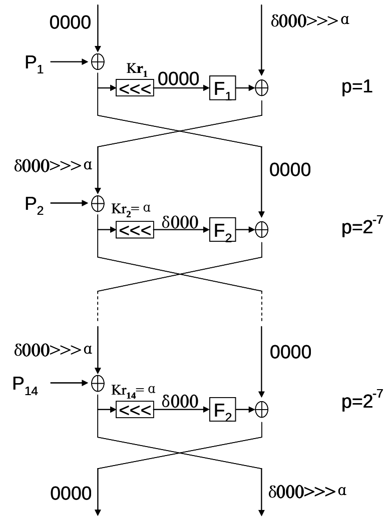


Fig 4: Differential characteristic for 14-round Blow-CAST-Fish

2. Loop for candidate α of kr_2 (α = 0, ……31):

(a) Collect all those plaintext pairs whose XOR is δ000>>> α at the right halves R_0. One structure proposes 2^{31} such pairs, thus 2^{20} structures propose 2^{51} such pairs. The number of right pairs is $2^{51}\times2^{-49}=2^2$.

(b) Loop for 2^{37} values of P_{18} and kr_{16} in round 16.

i. Loop for remaining pairs:

For all 2^{19} remaining ciphertext pairs

$$CL\oplus CL' = \delta000 >>> \alpha$$

calculate the outputxor of round function F_1 in the 16^{th} round. If the values of P_{18} and kr_{16} make the next equation holds, increase the corresponding key counter [$P_{18} \parallel kr_{16}$] by 1.

$F_1((CL\oplus P_{18})<<<kr_{16})\oplus F_1((CL'\oplus P_{18})<<<kr_{16})= CR\oplus CR'$
$=xyzt$

where C=(CL ‖ CR) and C'=(CL' ‖ CR') are a pair of ciphertexts. If the key counter is equal to 4, the subkey values P_{18}, kr_{16}, and kr_2 are the right subkey with high probability.

4.3 Complexity Analysis

As the analysis of 8-round, the signal to noise can be computed as follows,

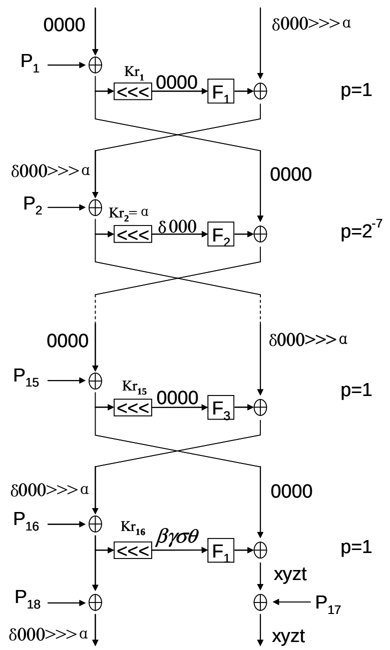


Figure 5: 16-round Blow-CAST-Fish

$$S/N = \frac{2^{-49} \times 2^{37}}{2^{5} \times 2^{-32}} = 2^{15}$$

where k=37,p=2^{-49},α=2^5,β=2^{-32} ,.The data complexity is 2^{52} chosen plaintexts. The time complexity is about $2^5 \times 2^{37} \times 2^{20} = 2^{62}$ times of one-round encryption which is equivalent to 2^{58} times of 16-round encryption. The attack is efficient for Blow-CAST-Fish with key sizes lager than 58 bits. And the proportion of weak keys is 2^{-52}.

5 Summary

Until now, there is no analysis result for the algorithm. In this paper, we first studied the properties of round function F_1 and F_3 with non-zero inputxor and zero outputxor in the 4th and 6th round of Blow-CAST-Fish separately, then 6-round differential characteristic with the probability 2^{-61} can be identified which are resulted by 2^{-12} of the total key space. Based on the 6-round differential characteristics, we can attack 8-round Blow-CAST-Fish.

Then we studied the properties of S_1 with non-zero inputxor and zero outputxor in round function of Blow-CAST-Fish, 14-round differential characteristic with the probability 2^{-49} can be identified which are resulted by 2^{-52} of the total key space. Based on the 14-round differential characteristics, we can attack 16-round Blow-CAST-Fish. We can also use S_2, S_3 or S_4 for analysis, the probability of weak keys which can produce collision with S_2 is nearly to the one with S_1. But the probability of weak keys which can produce collision with S_3 or S_4 is lower than that of S_1.

References

[1] Krishnamurthy G.N, Dr. V. Ramaswamy, Leela G.H and Ashalatha M.E, "Blow- CAST-Fish: A New 64-bit Block Cipher", Bapuji Institute of Engineering and Techology, Davangere-577004, Karnataka, India.

[2] B.Schneier, "Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish)",Fast Software Encryption, Cambirdge Security Workshop proceedings (December 1993), Springer-Verlag, 1994, pp.191-204.

[3] Adams, C. "The CAST-128 Encryption Algorithm". RFC 2144, May 1997.

[4] S. Vaudenay, "On the weak keys of Blowfish," Fast Software Encryption (FSE'96), LNCS 1039, D. Gollmann, Ed., Springer-Verlag, 1996, pp. 27–32.

[5] Seki. H and Kanekko. T, "Differential Cryptanalysis of CAST-256 Reduced to Nine Quad-rounds", IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, E84A(4), pp.913-918, 2001.

[6] J. Nakahara Jr and M. Rasmussen, "Linear Analysis of Reduced-round CAST-128 and CAST-256", SBSEG2007, pp.45-55, 2007.

[7] E. Biham and A. Shamir, "Differential Cryptanalysis of DES-like Cryptosystems", Journal of Cryptology, vol.4, no.1, pp.3-72, Springer, 1991.

[8] M. Q.Wang, X. Y.Wang, C. H. Hu, New Linear Cryptanalytic Results of Reduced-Round of CAST-128 and CAST-256, appeared on SAC'2008.

[9] M. Q.Wang, X. Y.Wang, K.P. Chow, Lucas C.K. Hui, Differential Crypanalysis of Reduced-Round of CAST-128, Preprint.

A Keyless Polyalphabetic Cipher

William J. Spezeski
Department of Computer Science
Massachusetts College of Liberal Arts
375 Church Street, North Adams MA 01247 USA
Phone: 413.662.5591
Email: w.spezeski@mcla.edu

***Abstract*-A polyalphabetic cipher is one in which multiple alphabets (monoalphabets) are used to encode a plaintext message. Each letter of plaintext is encoded by selecting one of the alphabets with which to determine a substitution. A key word or phrase is often used to select different alphabets to encrypt the plaintext. This paper considers three concepts that, when combined, lead to a keyless polyalphabetic cipher.**

INTRODUCTION – POLYPHABETIC CIPHERS

There are many variations of polyalphabetic cipher types with varying degrees of complexity. Most rely on a key or keyword to form the substance of the ciphertext. Consider this very simple example: A polyalphabetic cipher with the key BLUEBIRD would encode its first letter of plaintext using its B-alphabet, its second letter using its L-alphabet, its third letter using its U-alphabet and so on. The ninth letter thru the sixteenth letters of plaintext would reuse the B,L,U,E,B,I,R,D-alphabets for encipherment, respectively. Every eight letters of plaintext would use the key for encipherment until the plaintext message is exhausted. Both the sender and receiver would need seven distinct alphabets available designated by B, L, U, E, I, R, D, respectively. Some classic multiple-alphabet ciphers include the Vigenere Cipher [6], Gronsfeld Cipher [3], Porta Cipher [3] and the Enigma Cipher of WWII [4].

A KEYLESS POLYALPHABETIC CIPHER

The polyalphabetic cipher being considered here is the byproduct of combining three unrelated concepts. Consequently, we will look at the methodology of this cipher in three stages.

A. A Monoalphabet Substitution Cipher

To illustrate the concept behind a keyless alphabetic cipher, we first consider a simple alphabetic cipher which relies on the letters of the standard alphabet as shown in Figure1 below. To encode the word 'CIPHER', we embark on a 'search and count' process.

For illustration, we begin at position 0 (the letter 'A'); any pre-designated letter position will do.

The first letter of plaintext is obtained by counting the number of letters that we pass over as we go (left to right) from the starting position 'A' until we find the first plaintext letter 'C'. It is coded by the resulting count, in this case, 2.

The second letter of plaintext is obtained by counting the number of letters that we pass over as we go from the first plaintext letter 'C' to the second plaintext letter 'I'. The second letter is coded by the second count 6. And so on. If the search process takes us beyond $letter_{25}$ (Z), we return to $letter_0$ (A) and continue the count from right to left. Thus $letter_{26} = letter_0$ in our circular count.

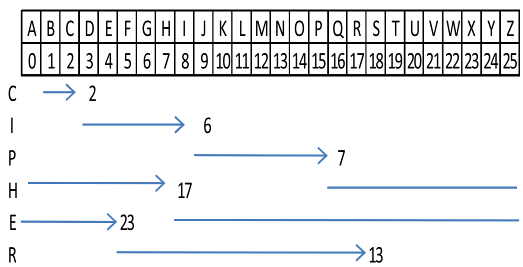


Fig. 1 The Search and Count Algorithm

Using this scheme, the plaintext word 'CIPHER' is encoded as the substitution ciphertext 2 6 7 17 23 13. This is the general idea that we shall use for substituting characters of plaintext with numbers to form the ciphertext. Note that it can be made more obtuse by setting up patterns of alternating clockwise and counter-clockwise counting.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*, DOI 10.1007/978-90-481-9151-2_92, © Springer Science+Business Media B.V. 2010

Some examples of various monoalphabets are shown below.

A1 = defghijklmnopqrstuvwxyzabc
A2 = jlpawiqbctrzydskegfxhuonvm
A3 = juliscaertvwxyzbdfghkmnopq
A4 = xzavoidbygerspcfhjklmnqtuw
A5 = fzbvkixaymeplsdhjorgnqcutw

B. A Substitution Cipher with a Polyalphabet

To develop our polyalphabet, we string several alphabets, say {A1, A2, A3, A4, A5}, together to form the polyalphabet. The same basic assignment method of converting plaintext to ciphertext shown above is used with the polyalphabet. A starting position must be identified; then the 'search and count' process begins. If the polyalphabet is a string of length N, then after reaching $letter_{N-1}$ in a count, the count continues with $letter_0$, $letter_1$, $letter_2$ and so on; thus $letter_N$ = $letter_0$.

A1	A2	A3	A4	A5

Fig. 2 – Polyalphabet Structure

The alphabets making up the polyalphabet need not be ordered or of the same length. The number of alphabets that constitute the polyalphabet may vary as well. Below is an example of a polyalphabet formed by concatenating alphabets A1, A2 and A3 above.

defghijklmnopqrstuvwxyzabcjlpawiqbctrzydskegfxhuonvm
juliscaertvwxyzbdfghkmnopq

In some sense, this polyalphabet can be thought of as a large monoalphabet. Because its construction (as I use it) makes use of several existing monoalphabets , I use the term polyalphabet to describe it.

C. A Keyless Polyalphabetic Cipher

Most polyalphabetic ciphers require a key - a word or phrase that distinguishes which particular monoalphabet a particular plaintext letter is being enciphered with. The classic Vignere Cipher [5] is a case in point. However, using the 'search and count' process mentioned above, there is no need for a key.

The 'search and count' process essentially looks for the first occurrence of the plaintext letter as it traverses the polyalphabet. A match results in a ciphertext number. However, there is no need to look for only first occurrences of a plaintext letter. We could search for the second occurrence of a plaintext letter, a third occurrence, and so on. Better yet, an arbitrary (random) number can determine what occurrence of each letter of plaintext we will search for in the polyalphabet. A computer can randomly generate these integers or they can be pulled off-the-top of the sender's head. The selection of these numbers results in different versions of the encoded plaintext, but does not affect the decoding of the ciphertext. The original plaintext and its corresponding decoded ciphertext are always the same. For example, each ciphertext below

Plaintext: h a p p y b i r t h d a y

Ciphertexts:

1. 36 75 16 55 10 57 57 35 64 30 16 78 52 52
2. 9 48 16 28 37 30 3 62 10 3 43 51 79 25
3. 63 75 43 55 37 57 3 8 64 57 16 78 25 79

is obtained by using the "Search and Count" algorithm with a standard 26- letter alphabet plus a 'space'. If we use five 27-character unsorted alphabets to construct a polyalphabet, some resulting ciphertexts for 'happy birthday' appear as follows:

1. 48 14 60 29 10 33 17 45 60 96 51 76 74 8
2. 48 37 12 54 56 14 9 26 60 27 11 63 24 67
3. 5 27 90 29 56 41 4 58 61 60 32 41 72 17
4. 74 86 56 17 81 69 4 58 31 43 79 41 79 60

All four ciphertexts above decrypt back to the original plaintext. Notice that the last series of plaintext ciphers consists of (with leading zeros) all 2-digit numbers with the largest number being 96, even though there are 135 characters in the polyalphabet that is being used. This is being controlled by the fact that the random number choices are restricted to 1, 2, or 3. This helps to disguise the number of 'alphabets' that form the polyalphabet as well as the length of the polyalphabet.

Decoding the Ciphertext

Deciphering the coded message is very easy. One starts at the pre-designated letter position of the same polyalphabet and simply counts, associating the final number of each count with the associated plaintext letter. Figure 1 above illustrates both how the word CIPHER can be encoded and how the ciphered word can be decoded. In the encipherment process, searching for a plaintext letter results in a count , which in turn becomes the ciphered substitute; in the decipherment process, the count becomes the means of locating the corresponding plaintext letter. As mentioned before, an enciphered text can have many possible variants. However, each variant will decipher to the same plaintext.

SOME OBSERVATIONS

Let us suppose that a five letter word is to be encoded with a 100-character polyalphabet using the random search 'modifiers' 1,2 and 3. Then there will be 100 possible starting positions and

$$100 \times 3 \times 3^2 \times 3^3 \times 3^4 \times 3^5 \quad \text{or} \quad 100 \times 3^{15}$$

possible variants of the enciphered five letter word. More generally, if a plaintext message consisting of N characters (including spaces) is encoded with a P-character polyalphabet using K random search modifiers, then the number of possible enciphered variants is

$$P\, K^{N(N+1)/2} .$$

If a 135-character polyalphabet to code 'happy birthday' using 3 random search modifiers, the number of variants would be 135×3^{105} . Four of the variants were shown above. For any fixed starting point, the number of variant ciphertexts for 'happy birthday' is 3^{105} .

IMPLEMENTATION

The implementation of the Search and Count Cipher (SCC) requires that the both the sender and receiver use the same polyalphabet as well as an agreed upon starting position within the polyalphabet. There are 26! (over 10^{26}) regular 26-letter monoalphabets available for building a polyalphabet. That number grows rapidly with an increase in alphabet size.

Polyalphabets can be easily constructed by having a collection of monoalphabets available for concatenation. The computer language Java, for example, provides string concatenation (+) that allows a wide variety of polyalphabets to be built very simply from existing monoalphabets. Again, this is a primary reason for referring to the cipher as a polyalphabetic cipher. In turn, multiple alphabets can be constructed from a single one by using a shuffle algorithm to rearrange an alphabet's letters. If both sender and receiver have the same set of numerically identifiable alphabets at hand, the sequence of alphabets used to construct the operable polyalphabet can be encoded within the numerical ciphertext. The starting position can be part of the ciphertext as well. And like the polyalphabet construction, the plaintext message string can be built by using concatenation. Coding the enciphering and deciphering algorithms in a computer language like Java is a straight-forward process using String arrays. Here are some things to keep in mind:

The number of alphabets used to form the polyalphabet should be a prime number. This is to ensure that the searches do not get confined to any cyclic subset of alphabets comprising the polyalphabet.

The number of distinct random numbers that get generated should be strictly less than the number of alphabets making up the polyalphabets. For example, assume that a particular polyalphabet consisted of 5 alphabets each having 25 characters. Assume further, that the distinct random numbers that get generated are 1, 2, 3. The largest possible numeric substitution for a plaintext letter would be 100, even though there are actually 125 characters in the polyalphabet. This helps to conceal the actual string length of the polyalphabet. In fact, the 5 alphabets mentioned can be rearranged so that only1- or 2-digit numbers will be generated for substitution. If the ciphertext is converted to ASCII characters , 3-digit numbers (< 256) can be accommodated without the ciphertext getting unwieldy.

Use the Java string method *length()* (or equivalent) to make any reference to the length of the active polyalphabet. It helps to further conceal its actual string length.

The alphabets that comprise the polyalphabet can vary in number, length, letter arrangement and order of appearance within the polyalphabet. The details, of course, must be agreed upon in advance by both sender and receiver. Below is the ciphertext that is generated from a polyalphabet formed by the 5 alphabets above (A1, A2, A3, A4, A5 each with an extra 'space') using modifiers 1, 2, 3 and starting at the first position of alphabet A1.

74 52 17 78 38 44 73 25 34 41 29 5 14 57 48 77 67 26 18 61 79 64 13 25 39 1 20 75 23 30 41 11 3 29 14 72 78 30 4 78 20 8 63 56 32 42 78 40 79 64 68 36 52 23 47 23 19 55 76 34 15 36 63 79 23 16 27 26 41 3 59 39 72 15 9 64 39 8 21 50 13 2 68 28 56 38 30 57 4 80 80 21 80 68 23 9 25 23 12 4 2 63 51 63 37 12 8 21 40 51 46 41 15

It is the ciphertext for the verse

speak roughly to your little boy
and beat him when he sneezes
he only does it to annoy
because he knows it teases

FIELD USE

This cipher system works very efficiently using a computer to encode and decode messages. It can also work very efficiently in the field. It would be helpful to have a cipher disk [4] or some graph paper and limit the number of alphabets making up the polyalphabets to 3, 5 or 7. The random integers can be picked at will without hindering the deciphering process. A drawback to SCC is that an error in counting will corrupt the rest of the ciphertext. A slide rule-like device could be very helpful to minimize this problem.

A polyalphabet constructed with five 30-character alphabets for field use would have 150 characters while a Vigenere table [4] will have 676 (26x26) characters.

CONCLUSION

In 1949, Claude Shannon [5], widely known as the Father of the Information Age, proposed the following characteristics of a good cipher:

1. The amount of secrecy needed should determine the amount of labor for encryption and decryption.
2. The set of keys and the enciphering algorithm should be 'free from complexity'.
3. The implementation of the process should be as simple as possible.
4. Errors in ciphering should not propagate and cause corruption of further information in the message.
5. The size of the encoded message should be no larger than the text of the original message.

The SCC produces random-like sequences of numbers for ciphertext that help to diffuse the process of cryptanalysis. The randomness of the ciphering algorithm reduces the recognizable statistical patterns that a cryptanalyst might use to a minimum. The lack of a key further frustrates cryptanalysis efforts.

In the previously encoded verse above, there are 108 integers in the ciphertext whose values fall in the interval [1, 80]. Figure 3 illustrates the frequency counts of numeric ciphertext in the interval [1, 80] and compares it with 108 randomly generated numbers in the same interval.

Frequency Count	Ciphertext	Random Sequence
0	21	22
1	29	26
2	15	18
3	9	10
4	4	4
5	0	0
6	1	0

Fig. 3 - Distribution of 108 Integers in the Interval [1, 80]

For this ciphertext, there were 21 values in [1, 80] that did not appear; There were 22 values in [1, 80] that did not appear in the random sequence. There were 29 values of ciphertext that had a frequency count of 1 while the random sequence had 26. And so on. The outlier is one value with frequency count 6.

Also of significance is that there is only one digram and no trigrams in this ciphertext. This, coupled with the randomness of the cipher and its keyless feature, helps to increase the effort of any analysis to break the code. All this speaks to Shannon's first requisite. That the SCC meets requisites two and three are clear. The SCC, however, does not meet the requirement four. An error in counting will compromise the rest of the ciphertext. On the other hand, computers are pretty reliable. The final requisite is not met either. The use of numbers for ciphertext will at least double the length of the message. If, on the other hand, the ciphertext is converted and sent as ASCII code characters, the ciphertext length will be roughly the same. Of course, it adds a step to the message encryption and decryption.

Will the SCC survive the test of cryptanalysis? That question remains.

REFERENCES

[1] Baldoni, Ciliberto & Cattaneo, *Elementary Number Theory, Cryptography and Codes*, Springer-Verlag, (2009).
[2] Friedman, William F., *Military Cryptanalysis - Part II: Simpler Varieties of Polyalphabetic Substitution Systems*, Filiquarian Publishing, LCC./Qontro (1938).
[3] Gaines, Helen F., *Cryptanalysis*, Dover Publications, (1956)
[4] Kahn, David, *The Code-breakers,* Macmillan Publishing, (1979)
[5] Shannon, Claude, "Communication Theory of Secrecy Systems", Bell System Technical Journal, (1949).
[6] Singh, Simon, *The Code Book*, Fourth Estate Ltd., (1999).
[7] Sinkov, Abraham, *Elementary Cryptanalysis,2nd edition*, The Mathematical Association of America, (2009).

Vertical Handover between WiFi Network and WiMAX Network According to IEEE 802.21 Standard

Mariem Thaalbi, Nabil Tabbane
Research Unit MEDIATRON
SUP'COM
Tunis, Tunisia
mariem.thaalbi@gmail.com nabil.tabbane@supcom.rnu.tn

Abstract—**IEEE 802.21 is one of the emerged standards designed to build handover between heterogeneous technologies while optimizing session continuity. In this paper we provide a performance evaluation of the vertical handover between WiFi and WiMAX using IEEE802.21 for different traffic flows: VoIP, video streaming and FTP. The performance metrics used for this evaluation include throughput, end-to-end delay and packet loss rate. The performance evaluation is done using the Network Simulator NS2.**

Keywords: WiFi; WiMAX; vertical handover; IEEE802.21; MIH

I. INTRODUCTION

The rapid expansion of heterogeneous wireless and mobile networks raises interesting issues such as vertical handover. To insure vertical handoff between WiFi and WiMAX networks, we will propose, in this paper, the IEEE 802.21 as an alternative to maintain uninterrupted user sessions during handovers between these two different networks. The IEEE 802.21 supports media Independent handover (MIH) in order to provide seamless mobility between heterogeneous networks. The purpose of this work is to evaluate the performances of IEEE 802.21 in vertical handover between WiFi and WiMAX technologies. The remainder of this paper is organized as follows: The background of WiFi and WiMAX are briefly viewed on Section II and III. An overview of IEEE 802.21 is given in Section IV. Simulation Scenario and experimental results are presented in Section V. The paper is concluded in Section VI.

II. WIFI NETWORKS

Wireless Fidelity (WiFi) [1] is a wireless local area network over limited range. It is also known under IEEE 802.11, and aims to connect devices such as Personal Computer, PDA, laptop, printers, etc.

IEEE 802.11 allows connectivity in infrastructure mode or ad hoc mode. In infrastructure mode, the WiFi stations communicate through the Access Point (AP). In this mode, the 802.11 network architecture, illustrated in Fig.1, is hierarchical. Its basic element is the Basic Service Set (BSS), which is the group of stations covered on the area covered by one AP. A BSS may also be part of larger network element called Extended Service Set (ESS). The ESS consists of one or more BSSs connected to the same wired LAN so called Distribution System (DS).

In ad hoc mode, devices can communicate directly with each other without Access Point support.

In our paper, we will focus on infrastructure mode.

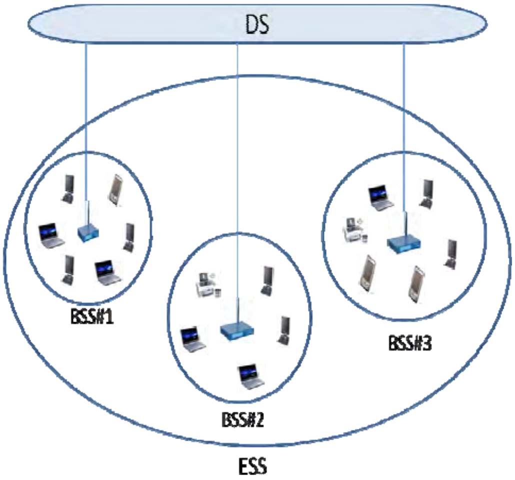


Fig.1. WiFi infrastructure mode.

III. WIMAX NETWORKS

Worldwide Interoperability for Microwave Access (WiMAX) [2] is a Wireless Metropolitan Access Network (WMAN). It is used generically to describe wireless systems based on IEEE 802.16.

WiMAX offers high data rates and large area coverage. It supports fixed and mobile broadband access. In our work, we will focus on Mobile WiMAX (IEEE 802.16e) [3].

The Mobile WiMAX Network, described in Fig.2, is divided into three parts: the Mobile Station (MS) which is the user's device, the Access Service Network (ASN) is the radio access and includes one or more Base Stations (BS) and one or more ASN Gateways (ASN-GW). The Connectivity Service Network (CSN) is the core of WiMAX Network. It provides IP connectivity to the MS.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_93, © Springer Science+Business Media B.V. 2010

In order to provide mobility, 802.16e supports handover mechanisms. The IEEE 802.16e defines three different types of handover: Hard Handover (HHO), Fast Base Station Switching (FBSS) and Macro Diversity Handover (MDHO). The HHO is often referred to as a break-before-make handover: first the MS disconnects from the serving BS and then connects to the target BS. In FBSS and MDHO, the MS maintains a diversity set which includes numerous active BSs in its range. The BS, to which the MS has the connection, is called the anchor BS. In FBSS, the MS transmits to and receives data from a single serving BS. The transition from the serving anchor BS to the target anchor BS in FBSS is done without invocation of the normal handover procedure, only the anchor BS update procedure is needed. In MDHO, the MS maintains a connection to one or more BSs in the diversity set simultaneously.

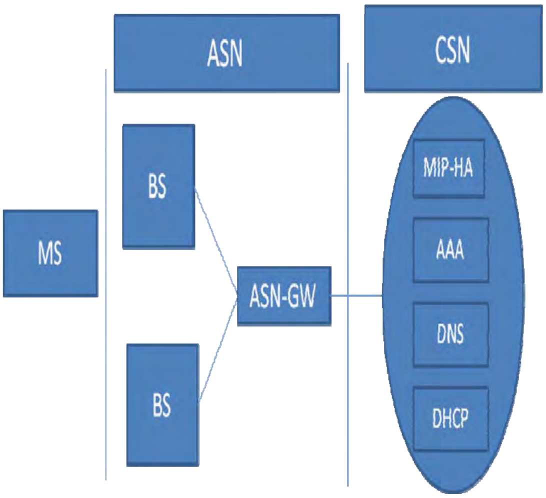


Fig.2. WiMAX architecture.

IV. 802.21 OVERVIEW

IEEE 802.21 [4] referred as Media Independent Handover (MIH) aims to enable seamless handover among heterogeneous networks. MIH defines a logical entity, Media Independent Handover Function (MIHF), located on layer 2.5 between link layer and network layer of the OSI model. It provides a framework that allows interaction between higher layers and lower layers. The MIHF supports three types of services: Media Independent Event Service (MIES), Media Independent Command Service (MICS), and Media Independent Information Service (MIIS). The MIES aims to provide and to predict link changes such as LINK_UP, LINK_DOWN, LINK_GOWING_DOWN, etc. These events are propagated from lower layers to upper layers through the MIH layer. MIES is divided into two categories, link events and MIH events. Link events are generated from the lower layer and transmitted to MIH layer. The MIH events are the events forwarded from MIH to upper layers. MICS refers to the commands, such as initiate handover and complete handover, sent from higher layers to lower layers. It allows enabling handover mechanism. MICS includes MIH command and Link command. MIH Commands originate from the upper layers down to the MIHF. Link Commands are specific to the lower layers. MIIS provides a framework by which MIHF can discover homogenous and heterogeneous network information existing within a geographical area to facilitate seamless handover when roaming across these networks. The MIIS provides a bidirectional way for the two layers to share information such as current QoS, performance information and availability of service. The figure bellow illustrates the MIH architecture:

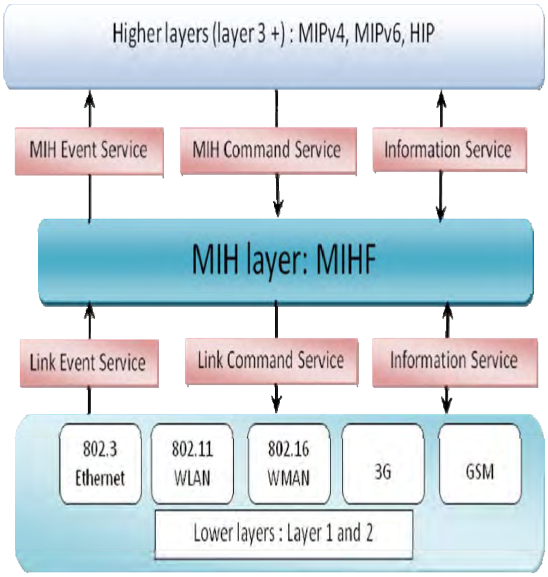


Fig.3. MIH architecture.

V. SIMULATION SCENARIO AND EXPERIMENTAL RESULTS

A. Simulation environement

The simulation results presented in this paper are obtained using Network Simulator NS-2 [5] with the MIH patch [6] and WiMAX patch [7].

The network simulator is used to define an appropriate topology that allows vertical handover between 802.16e cell and 802.11b cell. The WLAN and WiMAX parameters used during simulation are shown in TABLE I and TABLE II.

TABLE I. WLAN cell's parameters

IEEE 802.11b	
Coverage Raduis	100 m
Radio Propagation Model	Two- RayGround
Frequency	2.4 GHz
Transmission Power (Pt_)	0,0027 W
Receiving Threshold (RXThres)	2.64504e-10 W

TABLE II. WiMAX cell's parameters

IEEE 802.16e	
Coverage Raduis	1 Km
Radio Propagation Model	Two-RayGround
Frequency	3.5 GHz
Transmission Power (Pt_)	15 W
Receiving Threshold (RXThres)	**7.59375e-11**W
Sensitvity to link degradation (lgd_factor_)	**1.1**

B. Proposed scenario

The simulation topology, as shown in Fig.4, consists of one 802.11b AP, one 802.16e BS and one mobile station: MS. The MS is moving between the two cells with speed of 1 m/s, and then we increase the speed to 10 m/s.

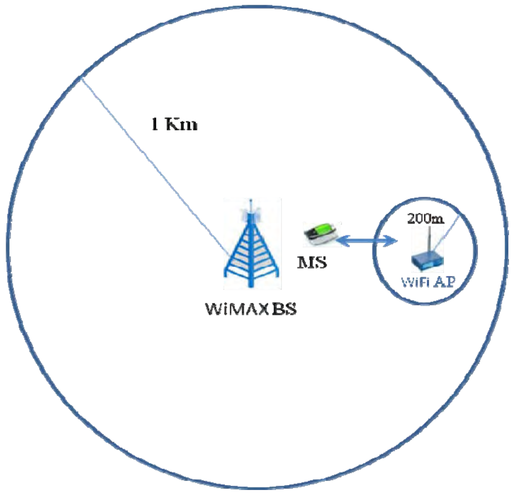


Fig.4. Simulation scenario

During simulations, the MS exchanges with its correspondent three types of traffic: VoIP, Streaming and File Transfer Protocol (FTP).The traffic flows parameters are shown on TABLE III.

TABLE III. Simulated traffic parameters

Traffic type	Packet size	Delay interval	Data rate
VoIP	160 bytes	20 ms	64 Kb/s
Streaming	2000 bytes	100 ms	160 Kb/s
FTP	1500 bytes	100 ms	120 Kb/s

The simulation duration was set to 250 s, the mobile node starts moving 10 s after the beginning of simulation time. The flow traffic starts from the beginning of the simulation.

C. Performance Criteria

In this paper, we focus on evaluating handover performance when the MS moves from WiMAX cell to WiFi cell and also the inverse direction from WiFi cell to WiMAX cell. The performance criteria adopted on our evaluation are: Throughput, end-to-end delay and packet loss ratio.

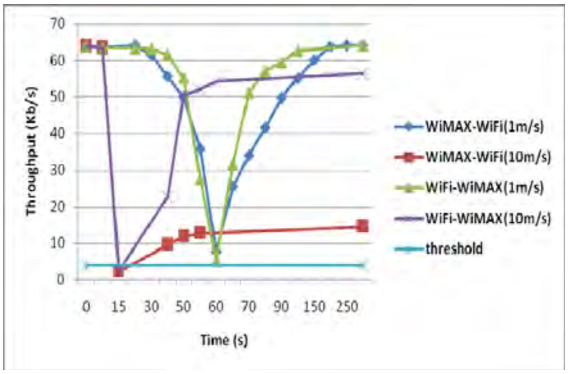
D. Results

Throughput

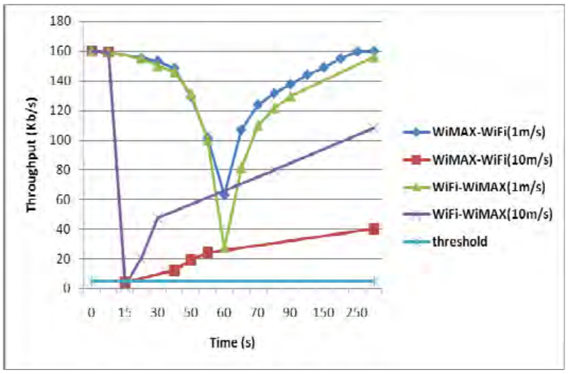
The throughput is the average rate of successful data delivery over a communication channel per unit time. It is measured in Kbits per second (Kb/s).

In this section we will calculate the throughput during simulation time for the two mobile speeds 1 m/s and 10 m/s with the two scenarios: handover from WiMAX to WiFi and from WiFi to WiMAX.

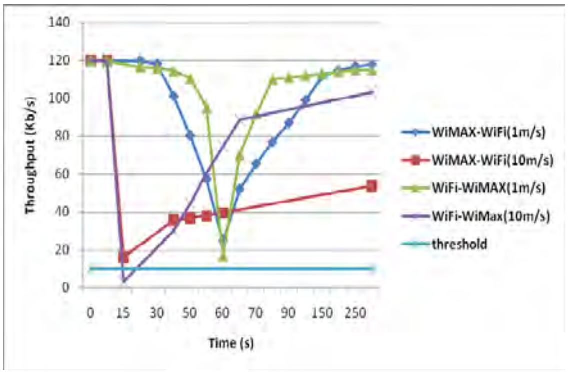
The throughputs during handover cases for different traffic are shown in Fig.5 (a) - (b) - (c).



(a) Throughput (Kb/s) - VoIP traffic



(b) Throughput (Kb/s) - Video streaming traffic



(c) Throughput (Kb/s) - FTP traffic

Fig.5. The throughput for different types of traffic

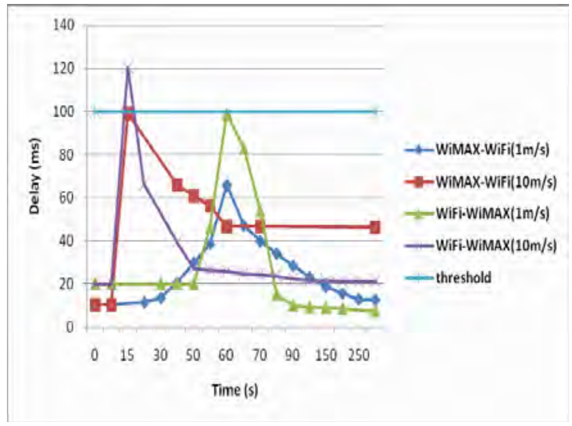
From these figures, we notice that the MS speed affects the effective throughput at handover moment. The thresholds fixed here to evaluate the throughput are equal to 4 Kb/s [8] for the VoIP traffic, 5 Kb/s [8] for the video streaming traffic and 10 kb/s [8] for the FTP traffic. With mobile speed equal to 1 m/s we obtain good results respecting the fixed thresholds. With MS speed fixed to 10 m/s the results obtained are not acceptable at handover moment. WiMAX network provides better throughput than WiFi network with mobility speed equal to 10 m/s.

End-to-end delay

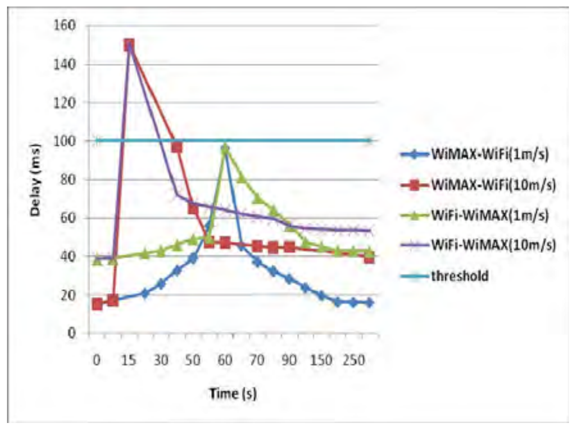
The end to end delay refers to the time taken by data packet to be transmitted across a network from source to destination. We calculate in this section the end-to-end delay with the same cases already described. The average delay for different simulated traffic is shown in Fig.6 (a) - (b) - (c).

When the MS speed is equal to 10 m/s, the delays degrade in WIFI network because it supports only the low mobility speed. The curves obtained with VoIP and FTP traffic during the handover process exceed slightly the threshold which is fixed to 100 ms for the VoIP and FTP traffics [8]. The curves obtained with streaming video traffic exceed very largely the threshold which is fixed to 100 ms [8].

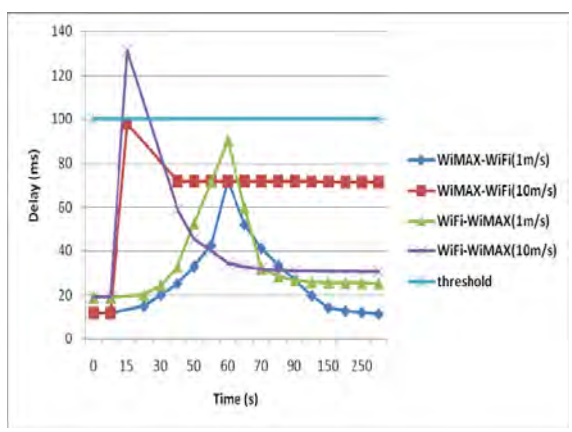
We note that, with the MS speed (10 m/s) the delays offered by WIMAX network are better than those offered by WIFI network because the mobile WIMAX network can support the low and the high mobility speed.



(a) End-to-end delay (ms) - VoIP traffic



(b) End-to-end delay (ms) - Video streaming traffic

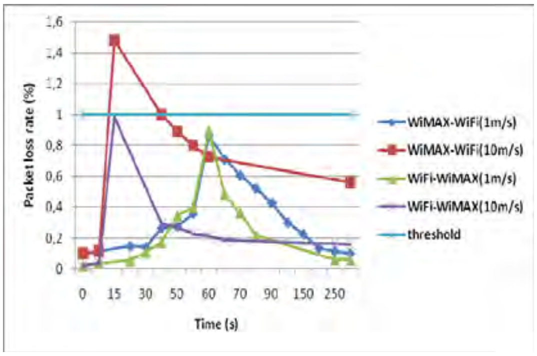


(c) End-to-end delay (ms) - FTP traffic

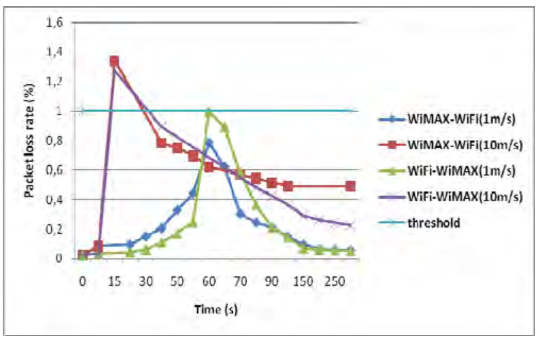
Fig.6. End-to-end delay for different types of traffic

Packet loss rate

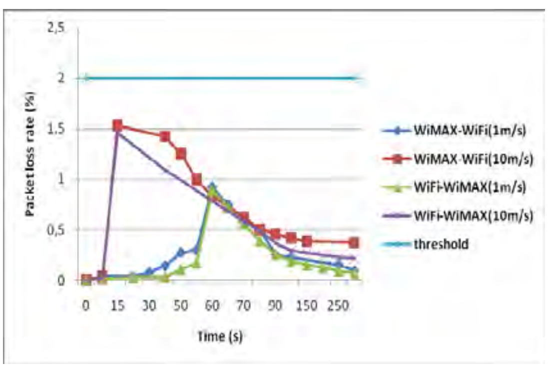
The packet loss rate parameter allows to determine the lost packet rate compared to those sent. The percentage of packet loss for different simulated traffics is illustrated in Fig.7 (a) - (b) - (c).



(a) Packet loss rate (%) - VoIP traffic



(b) Packet loss rate (%) - Video streaming traffic



(c) Packet loss rate (%) - FTP traffic
Fig.7. Packet loss rate for different type of traffic

In this figure, we see that with MS speed equal to 1 m/s, the two handover cases curves reflect good results, unlike the medium mobility speed equal to 10 m/s the handover curves exceed the threshold (fixed to 1%) for the VoIP and video streaming traffics [8].

With low mobility, the packet loss rate is less in WIFI network and the better results are obtained with the case when the MS makes a handover from WIMAX to WIFI. With MS speed equal to 10 m/s, the results are better in WIMAX network and the handover case from WIFI to WIMAX gives better packet loss rate values than the opposite case.

VI. CONCLUSION

The interoperability and the vertical handover between different networks present currently a real challenge to overcome.

In this paper, we have evaluated the performance of vertical handover between WiMAX and WiFi according to 802.21 for FTP and real-time traffic such as VoIP and video streaming. Seeing the results obtained, we can notice that the speed of the MS affects the handover performance. Also, the WiFi supports only the lox mobility speed. Finally, we can conclude that with the low speed the handover from WiMAX to WiFi generate best results than the opposite case of handover. But with medium speed, it is the otherwise because the 802.16e supports better the mobility management.

REFERENCES

[1] IEEE **Std 802.11™-2007**, IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, June 2007.

[2] IEEE Std 802.16**™-2004,** IEEE Standard for Local and Metropolitan Area Networks—Part 16: Air Interface for Fixed Broadband Wireless Access Systems, October 2004.

[3] IEEE Std 802.16e**™-2005**, IEEE Standard for Local and Metropolitan Area Networks—Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, February 2006.

[4] IEEE Std 802.21**™-2008**, IEEE Standard for Local and Metropolitan Area Networks- Part 21: Media Independent Handover, January 2009.

[5] Network Simulator- NS2- http://www.isi.edu/nsnam/ns, February 2009.

[6] National Institute of Standards and Technology (NIST), "The Network Simulator NS-2 NIST mobility package add-on,", March 2009.

[7] National Institute of Standards and Technology (NIST), "The Network Simulator NS-2 NIST IEEE 802.16 Implementation,", March 2009.

[8] Jeffery G. Andrews, Arunabha Gosh, Rias Muhamed, "Fundamentals of WiMAX, 1.7.3 Quality of Service", June 2007.

Study on the Impact of Node Density and Sink Location in WSN

Plamen Z. Zahariev, Georgi V. Hristov, Teodor B. Iliev
Department of Communication Systems and Technologies, University of Ruse, Ruse 7017, Bulgaria
E-mail: {pzahariev, ghristov, tiliev}@ccnrgroup.com

Abstract – **In this paper we analyze the total energy consumption of a wireless sensor network (WSN) per single communication round as a function of the node density and base station coordinates. We provide performance results of several simulations with WSN using both single hop and multi hop mechanisms for data transfer, and than discuss the relationships between the node density, the base station location and the total energy dissipation per round.**

Keywords – **WSN, Energy dissipation, Cluster, Cluster Head**

I. Introduction

Sensor networks play a major role in many aspects of society including home automation, consumer electronics, military application, agriculture, environmental monitoring, health monitoring and geophysical measurement. Usually sensor devices are small and inexpensive, so they can be produced and deployed in large numbers. Their resources of energy, memory, computational speed and bandwidth are severely constrained. Therefore, it is important to design sensor networks aiming to maximize their life expectancy. Different aspects of sensor networks such as data aggregation or fusion, packet size optimization, target localization, design challenges, network protocols are discussed in the literature with respect to crucial energy limitations and network lifetime maximization [1].

This paper presents a study focused on the total energy dissipation of a sensor network per a communication round. For this purpose we have simulated a WSN with a various number of sensors and with a base station, which coordinates are first at the center of the sensor field, than at the edge of the field and finally outside of the field.

The rest of the paper is organized as follows. In Section 2 we have provided the communication model, which is being used, together with the energy model. Section 3 presents the performance measures used in this paper. Section 4 provides the simulation scenarios and the various settings we use for the simulations. In Section 5 we provide the results of the conducted simulations as well as a discussion on the obtained results. Section 6 contains the conclusions on the conducted studies and directions for future work.

II. Model Overview

In this section we describe our model of a wireless sensor network with nodes, homogenous in their initial amount of energy and present the setting of the energy model. We consider a sensor network that is hierarchically clustered.

The LEACH (Low Energy Adaptive Clustering Hierarchy) protocol maintains such clustering hierarchy [2, 3]. In LEACH, the clusters are re-established in each round. New cluster heads are elected in each round and as a result the load is well distributed and balanced among the nodes of the network. Moreover each node transmits to the closest cluster head so as to split the communication cost to the sink [4]. Only the cluster head has to report to the base station and may expend a large amount of energy, but this happens periodically for each node. In LEACH there are an optimal predetermined percentage p_{opt} of nodes that has to become cluster heads in each round assuming uniform distribution of nodes in space. Since the nodes are homogeneous, which means that all the nodes in the field have the same initial energy, the LEACH protocol guarantees that everyone of them will become a cluster head exactly once every $1/p_{opt}$ rounds. Throughout this paper we refer to this number of rounds as epoch of the clustered sensor network. Initially each node can become a cluster head with a probability p_{opt} [5]. On average, $n \times p_{opt}$ nodes must become cluster heads per round per epoch. Nodes that are elected to be cluster heads in the current round can no longer become cluster heads in the same epoch. The non-elected nodes belong to the set G and in order to maintain a steady number of cluster heads per round, the probability of nodes $\in G$ to become a cluster head increases after each round in the same epoch. The decision is made at the beginning of each round by each node $s \in G$ independently choosing a random number in [0,1]. If the random number is less than a threshold $T(s)$ then the node becomes a cluster head in the current round. The threshold is set as:

$$T(s) = \begin{cases} \dfrac{p_{opt}}{1 - p_{opt}\left(r \bmod \dfrac{1}{p_{opt}}\right)} & if \quad s \in G \\ 0 & if \quad s \notin G \end{cases} \tag{1}$$

where r is the current round number. The election probability of nodes $\in G$ to become cluster heads increases in each round in the same epoch and becomes equal to one in the last round of the epoch. Note that by round we define a time interval where all clusters members have to transmit to the cluster head once. For the purpose of our study we use the energy model presented in Figure 1.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_94, © Springer Science+Business Media B.V. 2010

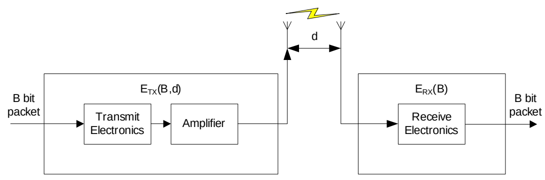


Fig.1 Radio Energy Dissipation Model

According to the radio energy dissipation model illustrated in Figure 1, in order to achieve an acceptable Signal-to-Noise Ratio (SNR) in transmitting a message with size of B bits, over a distance d, the energy expended by the radio is given by:

$$E_{TX}(B,d) = \begin{cases} B\times E_{elect} + B\times E_{fs}\times d^2 & if \quad d < d_o \\ B\times E_{elect} + B\times E_{ms}\times d^4 & if \quad d \geq d_o \end{cases} \quad (2)$$

where E_{elect} is the energy dissipated per bit of the transmitter or receiver, E_{fs} and E_{ms} depend on the transmitter amplifier model used and d is the distance between the sender and the receiver [6]. By equating the two expressions at $d = d_0$, we have $d_0 = \sqrt{\frac{E_{fs}}{E_{ms}}}$. The energy dissipated to receive a message with size B bits can be given by $E_{RX}(B) = B \times E_{elect}$ [7].

For our simulation studies we assume an area of 100×100 square meters with n number of nodes that are randomly distributed over the area. The energy dissipated in every cluster head node can be given using the following equation:

$$E^{CH} = B\times\left(E_{elect}\left(\frac{n}{k}-1\right)+E_{DA}\frac{n}{k}\right)+E_{TX}(B,d_{BS}), \quad (3)$$

where k is the number of cluster, E_{DA} is the energy consumption for processing of a bit per signal and d_{BS} is the distance between the cluster head and the base station [8]. The energy dissipation of a node in the wireless sensor network can be give using the following:

$$E^{Sensor} = E_{TX}(B,d_{CH}), \quad (4)$$

where d_{CH} is the distance between the sensor node and the cluster head. The total energy dissipation in the network per round can be given by:

$$E_{Total} = \sum_{i=1}^{n-k} E_i^{Sensor} + \sum_{j=1}^{k} E_j^{CH} \quad (5)$$

III. Performance Measures

We define here the measures used in this paper to evaluate the impact of the node density and base station location.

- *Node density* – since we assume a constant area of 100 x 100 square meters, this is the relationship between the number of nodes in this field and the surface of the field. Since the surface of the field is constant, higher node density means less distance between nodes [9].
- *Round* – the time interval for which all sensor nodes in a cluster transmit their packets to the cluster head.
- *Cluster head energy* – the energy dissipated by a cluster head for one round. This energy can be obtained using equation (3), and is equal to the sum of the energy for receiving B bits from all sensor nodes in the cluster, the energy for data aggregation of the received B bits from all sensor nodes in the cluster and the energy for transmitting B bit of information to the base station or a cluster head closer to the base station.
- *Sensor node energy* – the energy dissipated of a regular sensor in the cluster. The value of this measure can be obtained using equation (4), and is equal to the energy of transmitting B bits to the cluster head.
- *Total energy per round* – this measure is used to describe the whole energy dissipation by all devices in a wireless sensor network per round. The value of this measure can be obtained using equation (5).

IV. Simulation Scenarios

The goal of our study is to evaluate the impact of node density and base station location in wireless sensor networks. In order to complete our goal we have conducted several simulations with both single hop and multi hop scenarios. We have created two separate MatLab models, one for the single hop (Figure 2.) and one for the multi hop scenario (Figure 3.).

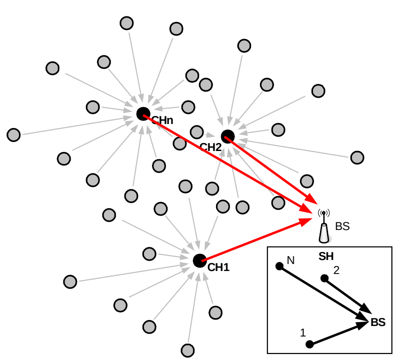


Fig.2 Single hop wireless sensor network

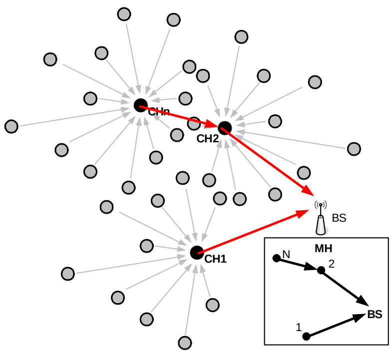


Fig.3 Multi hop wireless sensor network

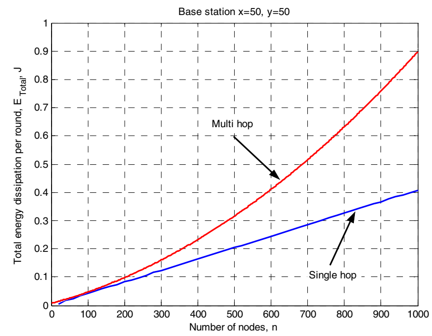


Fig.4 Total energy dissipation of the network, when the base station is in the center of the network

The area we use is 100×100 square meters and we randomly deploy n sensors in this area, where $20 \le n \le 1000$. For all simulations we assume that the nodes are homogenous in their initial amount of energy and that the size of the transmitted information is $B = 4000$ bits. Initially we place the base station at $x = 50$, $y = 50$ meters, which is in the center of the area. For the second scenario we move the base station at $x = 50$, $y = 100$ meters, which is at the edge of the area. For the last two scenarios the base station is outside the sensor network area at coordinates $x = 50$, $y = 200$ and $x = 50$, $y = 400$ meters, respectively. For all four scenarios we record the information about the performance measures of interest, shown above. The rest of the parameters used are given in Table 1.

TABLE I
SIMULATION PARAMETERS

Parameter	Description	Value
E_{elect}	Energy dissipation: electronics	50 nJ/bit
E_{fs}	Energy dissipation: amplifier	10 pJ/bit/m^2
E_{ms}	Energy dissipation: amplifier	13.10^{-4} pJ/bit/m^2
B	Transmit packet size	4000 Bits
n	Nodes number	$20 \le n \le 600$

V. SIMULATION RESULTS

On Figure 4 we present the results of the simulation, whit the base station in the center of the sensor network ($x = 50$, $y = 50$). It can be clearly seen that when the node density is relatively small ($n < 100$), the total energy consumption of the network for both multi hop and single hop scenarios is almost the same. This is due to the fact that the distance of every node in the field is smaller than d_0 and due to the fact that the number of clusters is relatively small, which means that there can be only few retransmissions between cluster heads. With the increase of the number of nodes the number of cluster also increases and the energy dissipation for the multi hop scenario increases greatly compared to the level of increase of the energy dissipation in the single hop scenario, due to the many retransmissions between cluster heads.

The simulation results obtained with the base station at the edge ($x = 50$, $y = 100$) of the sensor network are given on Figure 5. It is easy noticeable that the value of the total energy consumption for the single hop scenario is almost not affected by the repositioning of the base station. The reason for this is that only few cluster heads have distance to the base station larger than d_0.

The total energy dissipated in the multi hop scenario, on the other hand, is greatly affected by the repositioning of the base station. Analyzing the basic idea behind multi hop, we can come to the conclusion that, by moving the base station from the center to the edge of the network, the number of hops, for the data transmitted from the cluster heads close to the other edges of the network, will increase by approximately two times. This can be observed on Figure 5 and is the main reason for the quite larger value of the energy dissipated by the network devices in the multi hop scenario.

Figure 6 presents the results for the total energy dissipation as the base station is being relocated outside the wireless

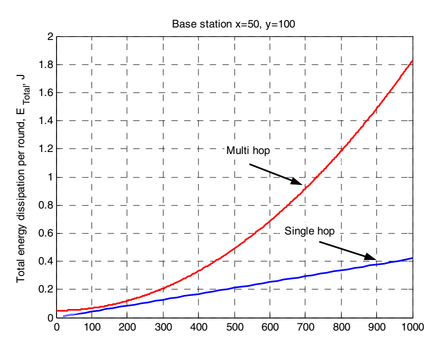


Fig.5 Total energy dissipation of the network, when the base station is at the edge of the network

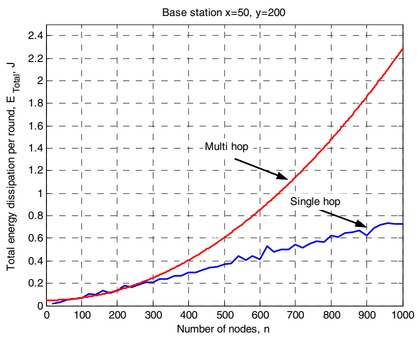


Fig.6 Total energy dissipation of the network, when the base station is at coordinates (x=50, y=200)

sensor network at x=50, y=200, which means that the distance of all sensors in the network to the base station will be $> d_0$. The results shown are quite interesting especially in the part where the density of the nodes is relatively small ($n < 100$). Unlike the previous results the multi hop scenario shows better results than the single hop. This can be explaned by the fact that the cluster number is relatively small and basically only few cluster heads will transmit to the base station (the other will transmit to a closer cluster head towards the base station), compared to the single hop scenario where all cluster heads will transmit to the base station. Again the situation changes with the increase of the number of sensors. The main difference is that both lines are closer than in the previous simulations. This is due to the fact that the larger transmission distance for the cluster heads in the single hop scenario is partially canceling the effect caused by the many retransmissions in the multi hop scenario.

Figure 7 presents the simulation results, obtained after the relocation of the base station even further away from the sensor network at x=50, y=400. It is obvious that the total energy dissipated in the single hop scenario is many times grater than the one in the multi hop scenario. This is due to the fact that in the multi hop scenario only few cluster heads

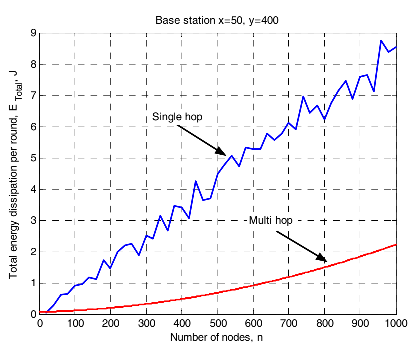


Fig.7 Total energy dissipation of the network, when the base station is at coordinates (x=50, y=400)

transmit to the base station compared to the single hop scenario, where all do. Compared to the results from the previous simulation here even the cancellation effect of the many hops that the data travels from the edge cluster heads to ones closer to the base station shows little effect, due to the larger distance to the base station.

VI. CONCLUSIONS AND FURTHER WORK

The contribution of this paper is a study of the impact of node density and base station location in wireless sensor networks. From the obtained simulation results it is clear to notice that with the relocation of the base station away from the sensor network area, the value of the energy dissipated in the singe hop scenario increases dramatically, compared to the multi hop scenario. The node density also has a critical impact on the total dissipated energy per round. This is especially evident in the multi hop scenarios shown on Figures 4, 5, and 6. A potential goal for further research can be the evaluation of node density and base station location in heterogeneous networks and design of a new approach for determining the range of the clusters and their optimal number.

ACKNOWLEDGMENT

This work is a part of the project BG051PO001/07/3.3-02/8–“MEQSIS”, funded by scheme “Support of the development of PhD students, postdoctoral, post-graduate and young scientists” from the program “Development of human resources” of the “European social fund”.

REFERENCES

[1] G. Smaragdakis, I. Matta, A. Bestavros, SEP: A stable Election Protocol for heterogeneous wireless sensor networks, In *Proceedings of International Workshop on SANPA*, 2004.

[2] W. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocol for wireless microsensor networks, In *Proceedings of the 33rd Hawaii International Conference on System Sciences* (HICSS-33), 2000.

[3] A. Chandrakasan, A. Smith, W. Heinzelman, An application-specific protocol architecture for wireless microsensor networks, In *IEEE Transactions on Wireless Communications*, vol. 1, pp. 660-670, 2002.

[4] I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, A survey on sensor networks. *IEEE Communications Magazine*, vol. 8, pp.102–114, August 2002.

[5] S. Bandyopadhyay, E. J. Coyle, An energy efficient hierarchical clustering algorithm for wireless sensor networks. In *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies* (Infocom 2003), April 2003.

[6] S. Bandyopadhyay and E. J. Coyle. Minimizing communication costs in hierarchically-clustered networks of wireless sensors, *Comput. Networks*, vol. 1, pp. 1–16, January 2004.

[7] C.-Y. Wan, A. T. Campbell, L. Krishnamurthy, PSFQ: A reliable transport protocol for wireless sensor networks, In *Proceedings of the 2nd IEEE Workshop on Applications and Services in Wireless Networks* (WSNA 2002), pp. 1–11, July 2002.

[8] C. Lin , M. Gerla, Adaptive clustering for mobile wireless networks, In *IEEE Journal on Selected Areas in Communications*, vol. 15 pp. 1265-1275, 1997.

[9] E. H. Callaway, *Wireless Sensor Networks, Architectures and Protocols*, Auerbach Publications, 2004.

An Algorithm for Synthesis of Families of Generalized Orthogonal Complementary Pairs

B. Y. Bedzhev, Member, IEEE
University of Shumen "Bishop Konstantin Preslavsky"
115 Universitetska Str.
Shumen, Bulgaria 9712, EU
M. P. Iliev
University of Ruse "Angel Kanchev"
8 Studentska Str.
Ruse, Bulgaria 7017, EU

***Abstract*-Families of generalized orthogonal complementary pairs (FGOCPs) are called families of pairs of radio signals, which autocorrelation functions (ACFs) and cross-correlation functions (CCFs) among every two pairs of a family have so-called zero correlation windows (ZCWs) in the area around the zero time shift. They allow the users to communicate practically without multiple access interferences in a multipath environment if the wireless communication systems works in a quasisynchronous mode. As a result the employment of the FGOCPs in the future wireless communication systems will significantly increase the data transmission rate. Due to this reason in the paper an algorithm for synthesis of FGOCPs is proposed. It distinguishes by the fact that it provides small level of the lobes of the CCFs without of the ZCW, which is a valuable feature.**

***Index Terms* – Synthesis of signals, Families of quasiorthogonal complementary pairs.**

I. INTRODUCTION

Today users and industry expect a significant enhancement of the quality of the services and the data transmission rate, offered by the wireless communication systems. This is a very hard technical challenge for the theorists and developers. From the information in the open literature [4], [5], [9] it can be concluded that a critical role for the whole performance of the wireless communication systems has the system of employed radio signals. Due to this reason great efforts are directed to finding of families of radio signals possessing both autocorrelation functions (ACFs) with a small level of the side-lobes and small cross-correlation functions (CCFs) among all pairs of members of a family. The focusing of the research activities on this problem can be explained by the following facts [1], [2], [4], [5], [6], [9], [10], [11], [13], [14], [15], [16], [17]. First, the pointed out families of radio signals allow the so-named self-interference (SI), caused by multipath spreading of electromagnetic waves, to be reduced by a separate processing of the direct and reflected signals. Second, it is possible the negative effect of simultaneous transmission of numerous users, named multi user interference (MAI), to be minimized.

A promising candidate for employment in the future wireless communication systems are the so – named *Families of generalized orthogonal complementary pairs* (*FGOCPs*). They are families of pairs of radio signals, which ACFs and CCFs among every two pairs of a family have so-called zero correlation windows (ZCWs) in the area around the zero time shift [4], [5], [6]. They allow the users to communicate practically without MAI in a multipath environment if the wireless communication systems work in a quasisynchronous mode. It should be mentioned that the so – named *families of orthogonal complementary codes* (*FOCCs*) provides possibility the users to communicate practically without any SI and MAI in a truly asynchronous mode [1], [4]. Anyway, in a comparison with FOCCs valuable advantages of the FGOCPs are the simplicity of their technical realization and the higher resistance to the frequency selective fading.

With regard to above described situation, our paper aims to develop an algorithm for synthesis of FGOCPs, which distinguishes by the fact that it provides small level of the lobes of the CCFs without of the ZCW. The importance of the studied problem results from the following facts. First, today it is possible to provide a quasisynchronous mode of operation of the wireless communication systems by a real time measurement of the location of the users relatively to the base stations of the cells. Second, despite of taken efforts many problems in the field of the synthesis of the FGOCPs are open still [4], [5], [6].

Paper is organized as follows. First, the basics of the FGOCPs are recalled. Second, our algorithm for synthesis of FGOCPs is suggested. At the end some important conclusions are given.

II. BASICS OF THE SYNTHESIS OF THE FAMILIES OF GENERALIZED ORTHOGONAL COMPLEMENTARY PAIRS

Let $K \geq 2, N \geq 1$, be integers and let us consider a family $W(K, 2, N, Z_0)$, comprising K pairs as every pair consists of two sequences with (code) length N:

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_95, © Springer Science+Business Media B.V. 2010

$$\{a_k(m,n)\}_{n=0}^{N-1} = \\ = \{a_k(m,0), a_k(m,1), ..., a_k(m,N-1)\}, m = 0,1 \quad (1)$$

Then $W(K, 2, N, Z_0)$ is called family of generalized orthogonal complementary pairs (FGOCPs) if the aggregated ACFs of the sequences of every pair and the aggregated CCFs of the sequences of every two pairs have so-called zero correlation windows (ZCWs) in the area around the zero time shift [4], [5], [6].

These conditions can be mathematically described as follow:

$$R_{A_k,A_l}(r) = \begin{cases} 2.N, \ k = l \cap r = 0; \\ 0, \ k \neq l \cap r = 0; \\ 0, \ 0 < |r| \leq Z_0, \end{cases} \quad (2)$$

where:

- r is the time shift, $r = -(N-1), (N-2), ..., -1, 0, 1, ..., N-1$;
- Z_0 is the width of the ZCW;
- A_k, $k = 0,1,...,K-1$ is the matrix, denoting the k-th pair of the family $W(K, 2, N)$, i.e.:

$$A_k = \begin{bmatrix} a_k(0,0) & a_k(0,1) & ... & a_k(0,N-1) \\ a_k(1,0) & a_k(1,1) & ... & a_k(1,N-1) \end{bmatrix} \quad (3)$$

- $R_{A_k,A_l}(r)$ is the aggregated CCF (ACF if $k = l$) of the rows of the k-th and l-th pairs (matrices), $k,l = 0,1,...,K-1$.

The aggregated $R_{A_k,A_l}(r)$ CCF (ACF if $k = l$) is evaluated by the well known formula [1], [2], [3], [4], [7], [8], [9], [15], [16]:

$$R_{A_k,A_l}(r) = \\ = \begin{cases} \sum_{m=0}^{1}\sum_{n=0}^{N-1-|r|} a_k^*(m,n).a_l(m,n+|r|), \ -(N-1) \leq r \leq 0, \\ \sum_{m=0}^{1}\sum_{n=0}^{N-1-r} a_k(m,n).a_l^*(m,n+r), \ 0 \leq r \leq N-1. \end{cases} \quad (4)$$

Here the symbol "*" means "complex conjugation".

It should be mentioned that FGOCPs are mathematical models of the radio signals for the code-division multiple access (CDMA) wireless communication systems working in quasisynchronous mode. In such a system K user can simultaneously communicate practically without SI and MAI if the following conditions are available. The k-th set (i.e. the matrix A_k, $k = 0,1,...,K-1$) of the exploited FGOCP is assigned to the k-th user as its signature code. The entry $a_k(m,n)$, $m = 0,1,...,M-1$, $n = 0,1,...,N-1$ of the matrix A_k, $k = 0,1,...,K-1$ is the complex envelope of the n-th elementary pulse of the m-th sequence. In fact the m-th sequence describes mathematically the law of the amplitude and phase modulation of the m-th subcarrier frequency during the every clock interval in the process of the spreading of the spectrum of the radio signals. Most often only phase manipulation (Q-phase shift keying) is used, i.e.

$$\forall a_k(m,n) \in \{\exp[(2\pi i l)/Q]; l = 0,1,...,Q-1\}, \ i = \sqrt{-1}, \quad (5)$$

but this restriction will not be used in the sequel.

It should be pointed out that for a FGOCP the number of the subcarrier frequencies is $M = 2$. Due to this reason, the both signals of every complementary pair can be simultaneously conveyed to the receivers by means of so – named quadrature modulation [4]. This peculiarity is a very valuable feature of the FGOCPs, because it makes possible to use only one carrier frequency with two orthogonal phase shifts (0 and $\pi/2$ respectively) instead of two subcarrier frequencies. As a result, the technical complexity of the wireless communication system and the negative influence of the so – named Doppler shift of the subcarrier frequencies, caused by the moving of the users, can be significantly reduced. Consequently, from a philosophical point of view the FGOCPs combine the advantages of the orthogonal complementary codes [1], [4] and the simplicity of the existing direct spreading (DS) CDMA wireless communication systems [4], [5], [6], [9]. All these arguments explain the interest to the methods for synthesis of FGOCPs.

For brevity but without loss of generality we shall suppose that $0 \leq r \leq N-1$ in (3). Then, after changing the order of summations, (4) can be transformed as follows:

$$R_{A_k,A_l}(r) = \sum_{n=0}^{N-1-r}\sum_{m=0}^{1} a_k(m,n).a_l^*(m,n+r) = \\ = \sum_{n=0}^{N-1-r} CA_k(n) \otimes CA_l^*(n+r). \quad (6)$$

Here $CA_k(n)$ is the n-th vector – column of the matrix A_k and the symbol matrix "$\otimes$" denotes the inner product of the n-th vector – column of the matrix A_k and the complex conjugated (n+r)-th vector – column of the matrix A_l. Now it should be recalled that the evaluation of the CCFs can be simplified by means of multiplication of polynomials [3]. More specifically, let us introduce the following polynomials

$$F_k(x) = \sum_{n=0}^{N-1} CA_k(n).x^n, \ k = 0,1,...,K-1, \quad (7)$$

$$F_l^*(x^{-1}) = \sum_{j=0}^{N-1} CA_l^*(j).x^{-j}\ ,\ l = 0,1,...,K-1\ , \tag{8}$$

then the coefficients of their polynomial product $F_k(x).F_l^*(x^{-1})$ will be the aggregated cross-correlations $R_{A_k,A_l}(r)$ for all time shifts $r = -(N-1), -(N-2), ..., -1, 0, 1, ..., N-1$ according to (6), i.e.:

$$F_k(x).F_l^*(x^{-1}) = \sum_{r=-(n-1)}^{N-1} R_{A_k,A_l}(r).x^r \tag{9}$$

It should be stressed that multiplication of the coefficients in the left side of (9) is performed as multiplication of vectors, i.e.

$$CA_k(n) \otimes CA_l^*(n+r) = \sum_{m=0}^{1} a_k(m,n).a_l^*(m,n+r) \tag{10}$$

and due to this reason it is possible $CA_k(n) \otimes CA_l^*(n+r) = 0$. Besides, the coefficients $R_{A_k,A_l}(r)$ of the right side of (9) satisfy the conditions (2).

III. An Algorithm for Synthesis of Families of Generalized Orthogonal Complementary Pairs

The complementary codes were invented by M. Golay in 1949. In his origin works [7], [8] K = 1, M = 2 and $\forall a(m,n) \in \{-1,+1\}$, $m = 0,1$, $n = 0,1,...,N-1$, which gave the reason these codes to be named complementary pairs. Later, in 1972, Tseng and Liu [16] extended the Golay's conception and introduced the so - named mutually orthogonal complementary sets, consisting of $M \geq 2$ complementary sequences.

Due to their unique correlation properties the complementary codes have been intensively studied. For instance the study in [4] is based on over 300 journal papers and conference reports, concerning this theme. The complementary codes have also exploited in some types radars and specialized communication systems and obtained experience has proved the practical effectiveness of these signals. As a result at the beginning of the new century it was suggested the FGOCPs, described in the previous section, to be the one of the base tools for multiple access to the communication channel in the next generations wireless communication systems [4], [5], [6], [9].

At the moment the used algorithms for synthesis of families of signals with ZCWs have the following shortcomings.

First, very often these algorithms generate only periodic signals with desirable correlation properties [5], [6], [9]. It should be stressed that in the real wireless communication systems the aperiodic correlation properties of the exploited signals are much more important [4].

Second, a big part of known algorithms are based on specific techniques [4] so that it is hard or impossible to be unified.

With regard to above shortcomings in this part of the paper an algorithm for synthesis of FGOCPs will be suggested. It generates FGOCP, which aperiodic correlation properties not only meet the conditions (2), but also provides small level of the correlation lobes for all non-zero time shifts. Besides our algorithm allows a significant part of known today techniques for synthesis of FGOCPs to be viewed from a common theoretical base.

The algorithm for synthesis of FGOCPs, presented in the sequel, is based on the following proposition.

Proposition 3.1: Let A be a set of M = 2 complementary sequences with length N and let D be a matrix

$$D = \begin{bmatrix} d(0,0) & d(0,1) & ... & d(0,K-1) \\ d(1,0) & d(1,1) & ... & d(1,K-1) \\ ... & ... & ... & ... \\ d(K-1,0) & d(K-1,1) & ... & d(K-1,K-1) \end{bmatrix} \tag{11}$$

which columns are mutually orthogonal, i.e.

$$\sum_{k=0}^{K-1} d(k,n).d^*(k,l) = \begin{cases} K, & n = l; \\ 0, & n \neq l, \end{cases} \tag{12}$$

for $0 \leq n, l \leq K-1$. Then the row - submatrices of the tensor product

$$E = \begin{bmatrix} d(0,0).A & d(0,1).A & ... & d(0,K-1).A \\ d(1,0).A & d(1,1).A & ... & d(1,K-1).A \\ ... & ... & ... & ... \\ d(K-1,0).A & d(K-1,1).A & ... & d(K-1,K-1).A \end{bmatrix} \tag{13}$$

are a FGOCP with parameters: family size – K, length of the sequences – $N' = K.N$, width of the ZCW – $Z_0 = N$.

Proof: According to (7) and (8) let $F(x)$ and $F^*(x^{-1})$ be the polynomials of (N-1)-th degree which coefficients are the vector – columns of the matrix A. As A is a set of M = 2 complementary sequences with length N ([4], [7], [8]):

$$F(x).F^*(x^{-1}) = M.N = 2.N\ . \tag{14}$$

Analogously, let $H_m(x)$ and $H_l^*(x^{-1})$ be the polynomials of (K-1)-th degree which coefficients are the entries of the m-th and l-th rows of the matrix D respectively:

$$H_m(x) = \sum_{n=0}^{N-1} d(m,n).x^n,\ m = 0,1,...,K-1,$$
$$H_l^*(x^{-1}) = \sum_{j=0}^{N-1} d^*(l,j).x^{-j},\ l = 0,1,...,K-1 \quad (15)$$

Using (15), the CCF of the m-th and l-th rows of the matrix D can be presented in the following polynomial form:

$$P_{H_m,H_l}(x) = H_m(x).H_l^*(x^{-1}) = \sum_{r=-(k-1)}^{K-1} R_{H_m,H_l}(r).x^r\ . \quad (16)$$

It is not hard to prove that the rows of the square matrix D are mutually orthogonal as a consequence of (12). Due to this reason

$$R_{H_m,H_l}(0) = 0\ . \quad (17)$$

Let $RE_m(k)$ be the m-th row – submatrix of the tensor product E:

$$RE_m(k) = [d(m,0).A \quad d(m,1).A \quad ... \quad d(m,K-1).A]. \quad (18)$$

After taking into account (14) - (18), the CCF of the m-th and l-th row - submatrices $RE_m(k)$ and $RE_l(k)$ of the tensor product (13) can be evaluated as follows

$$\left[\sum_{k=0}^{K-1} d(m,k).F(x).x^{k.N}\right]\left[\sum_{j=0}^{K-1} d^*(l,j).F^*(x^{-1})x^{-j.N}\right] =$$
$$= F(x).F^*(x^{-1})\left[\sum_{k=0}^{K-1} d(m,k).x^{k.N}\right]\left[\sum_{j=0}^{K-1} d^*(l,j).x^{-j.N}\right] \quad (19)$$
$$= 2N.H_m(x^N).H_l^*(x^{-N}) = 2.N.P_{H_m,H_l}(x^N).$$

Equation (19) proves the Proposition 3.1, because the coefficients of the polynomial in the right side of (19) satisfy the following restrictions

$$2.N.R_{H_m,H_l}(s) = \begin{cases} 0,\ s \neq r.N; \\ 2.N.R_{H_m,H_l}(r), s = r.N, \end{cases} \quad (20)$$
$$s = -(K-1).N, -(K-1).N+1, ..., -1, 0, 1, ..., (K-1).N$$

and

$$R_{H_m,H_l}(r) = \begin{cases} K, r = 0 \cap l = m; \\ 0, r = 0 \cap l \neq m, \end{cases} \quad (21)$$

as the rows of the matrix D are mutually orthogonal.

On the base of the Proposition 3.1, FGOCPs can be synthesized by the following Algorithm.

Algorithm 3.2:

1) Choose an initial complementary pair A:

$$A = \begin{bmatrix} a(0,0) & a(0,1) & ... & a(0,N-1) \\ a(1,0) & a(1,1) & ... & a(1,N-1) \end{bmatrix} \quad (22)$$

so that

$$N = Z_0,$$
$$a(m,n) \in \{\exp[(2\pi il)/q]; l = 0,1,...,q-1\},\ i = \sqrt{-1}, \quad (23)$$
$$m = 0,1,\ n = 0,1,...,N-1$$

In (23) Z_0 is the desirable width of the ZCW, q should be a divisor of Q, where Q denotes the type of phase manipulation, which will be used in the developed wireless communication system.

2) Choose a square orthogonal matrix D, which size is equivalent to the desirable number K of the users of the developed wireless communication system. Besides, the entries of the matrix D must belong to the set

$$d(m,n) \in \{\exp[(2\pi il)/s]; l = 0,1,...,s-1\},\ i = \sqrt{-1},$$
$$m = 0,1,...,K-1,\ n = 0,1,...,K-1, \quad (24)$$

where s is a divisor of Q.

It should be mentioned that in the open literature detailed tables with complementary pairs and orthogonal matrices are given [4], [8].

3) Form the tensor product E, according to (13).

The rows of E are a FGOCP with parameters: family size – K, length of the sequences – $N' = K.N$, width of the ZCW – $Z_0 = N$.

IV. RESULTS

It should be stressed that the above Algorithm 3.2 gives the valuable opportunity to minimize correlation lobes without of the ZCW of the synthesized FGOCPs by an appropriate selection of the matrix D. With regard to the known upper bounds of the level of the lobes of the CCFs of the binary sequences [12], an exhaustive examination of the correlation properties of the square orthogonal matrices have been conducted. As a result we have found that the most appropriate matrices D are these, which are generated by the so – named *Frank's method* [2]. This conclusion issues from the following propositions.

Proposition 4.1: Let g be a K-th primitive root of the unity:

$$g = \exp[(2\pi il)/K],\ i = \sqrt{-1} \quad (25)$$

where l and K are relatively prime numbers.

Then the square matrix, obtained by the Frank's method:

$$G=\begin{bmatrix} g^{(0.0)} & g^{(0.1)} & \dots & g^{(0.(K-1))} \\ g^{(1.0)} & g^{(1.1)} & \dots & g^{(1.(K-1))} \\ \dots & \dots & \dots & \dots \\ g^{((K-1).0)} & g^{((K-1).1)} & \dots & g^{((K-1).(K-1))} \end{bmatrix} \qquad (26)$$

has cyclically orthogonal rows (and columns) for all shifts.

Proof: Let us consider the periodic CCF of the *n*-th and *l*-th columns of the matrix (26) for an arbitrary shift $r=0,1,\dots,K-1$

$$\begin{aligned} P_{CG_n,CG_l}(r) &= \sum_{k=0}^{K-1} g^{k.n}.g^{-(k+r).l} = \\ &= g^{-l.r}\sum_{k=0}^{K-1} g^{(n-l).k} = \begin{cases} K.g^{-n.r},\ n=l; \\ 0,\ n\neq l. \end{cases} \end{aligned} \qquad (27)$$

Equation (27) proves the Proposition 4.1.

As the rows (and columns) of the matrix (26) are cyclically orthogonal for all shifts, the matrix *G* satisfies the condition (12). Consequently, we can use the matrix *D* = *G* in the Algorithm 3.2, in order to synthesize a FGOCP.

Proposition 4.2: The lobes of the aperiodic CCFs of the columns (rows) of the matrix *G* do not exceed $K/2$.

Proof: From (4) and (27) follows that the periodic CCF of the *n*-th and *l*-th columns of the matrix (26) is

$$P_{CG_n,CG_l}(r) = R_{CG_n,CG_l}(r) + R_{CG_n,CG_l}(K-r). \qquad (28)$$

Here $R_{CG_n,CG_l}(r)$ and $R_{CG_n,CG_l}(K-r)$ are the aperiodic CCFs of the *n*-th and *l*-th columns of the matrix *G* for shifts *r* and *K-r* respectively. Without loss of generality let us suppose that

$$r \le K-r. \qquad (29)$$

Consequently

$$r \le \frac{K}{2}. \qquad (30)$$

From (27) and (28) follows:

$$\left|-R_{CG_n,CG_l}(r)\right| = \left|R_{CG_n,CG_l}(K-r)\right|. \qquad (31)$$

The magnitude of the right side of (31) can be estimated from the chain of inequalities:

$$\begin{aligned} \left|R_{CG_n,CG_l}(K-r)\right| &= \left|\sum_{k=0}^{K-1-(K-r)} g^{k.n}.g^{-(k+r).l}\right| \le \\ &\le \sum_{k=0}^{r-1}\left|g^{k.n}.g^{-(k+r).l}\right| = r \le \frac{K}{2} \end{aligned}. \qquad (32)$$

This proves the Proposition 4.2.

In fact, we have found that the aperiodic ACFs and CCFs of the rows (columns) of the matrix (26) have side – lobes which are smaller than the upper bound, found in the Preposition 4.2. Namely, for large number of values, applicable in the practice, the maximum side – lobes do not exceed $\sqrt{K}$.

This conclusion will be explained by the following example.

Example 4.3: Let *K* = 4. Then the matrix (24) has the form:

$$G=\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix},\ i=\sqrt{-1} \qquad (33)$$

The lobes of the aperiodic CCFs of the rows of the matrix (33) are given in the following table:

TABLE I
LOBES OF CCFS OF THE ROWS OF THE MATRIX (33)

r	-3	-2	-1	0	1	2	3
$R_{1,2}$	1	1+i	i	0	-1	-1-i	i
$R_{1,3}$	1	0	1	0	1	0	1
$R_{1,4}$	1	1-i	i	0	-1	-1+i	i
$R_{2,3}$	i	-1-i	1	0	-i	1+i	-1
$R_{2,4}$	i	0	i	0	-i	0	i
$R_{3,4}$	-1	1+i	-i	0	1	-1-i	i

In Table I *r* is the time shift and $R_{k,l}$ is the CCF of the *k*-th and *l*-th row of the matrix (33).

V. CONCLUSION

In the paper an algorithm for synthesis of FGOCPs is suggested on the base of Proposition 3.1. It allows the users to communicate practically without MAI in a multipath environment if the wireless communication systems work in a quasisynchronous mode. It should be stressed that a quasisynchronous mode can be provided by a measurement of the locations of the users respectively to the base station and introducing a time schedule for sending messages to the base station. This is not a hard technical task and as a result it is possible to exploit quasisynchronous mode in the next generations wireless communication systems.

The FOCCs provides possibility the users to communicate practically without any SI and MAI in a truly asynchronous mode

[1], [4], [5], [6]. Anyway, in a comparison with FOCCs valuable advantages of the FGOCPs are the simplicity of their technical realization and the higher resistance to the frequency selective fading.

The algorithm for synthesis of FGOCPs, suggested in the paper, has the following positive features.

First, it gives a general approach to the synthesis of the FGOCPs. Namely it is a direct construction for generating FGOCPs, providing both the desirable number of users K and the width Z_0 of the ZCW.

Second, in contrast with large number of constructions [5], [6] it is developed on the base of the aperiodic correlations which is the typical case for the present wireless communication systems.

Third, it allows minimizing of the correlation lobes without of the ZCW of the synthesized FGOCPs by an appropriate selection of the matrix D. For example, if we use matrix $D = G$ in Algorithm 3.2, as G is generated by the so – named *Frank's method* [2] (i.e. by (26)), then the maximum side – lobes do not exceed $\sqrt{K}.N$. Here K is the size of the matrix D, determining the number of the users of the developed wireless communication system, N is the length of the used pair of complementary sequences and $K.N$ is the amplitude of the central peak of the aperiodic ACFs of the FGOCP, synthesized by the Algorithm 3.2.

With regard to all above stated we hope the algorithm for synthesis of FGOCPs, suggested in the paper, to be useful in the process of development of the new generation wireless communication system, which aim to provide a radical enhancement of the data transmission rate.

ACKNOWLEDGMENT

The authors wish to thank Prof. Dr Kunchev, Prof. E. Ferdinandov, Prof. Dr H. Hristov and Prof. Dr B. Tsankov. Finally, the authors appreciate the thoughtful critique by the referees.

REFERENCES

[1] B. Y. Bedzhev and M. P. Iliev, "A general method for synthesis of families of orthogonal complementary codes," unpublished

[2] J. J. Benedetto and J. J. Donatelli, "Ambiguity function and frame-theoretic properties of periodic zero-autocorrelation waveforms," *IEEE J.of Selected Topics in Signal Processing*, vol.1, No.1, pp. 6-20, June 2007.

[3] R .E. Blahut, *Theory and Practice of Error Control Coding*, Addison-Wesley, Reading, Mass., USA, 1987

[4] H.-H. Chen, *Next Generation CDMA Technologies*, John Wiley & Sons, Chichester, West Sussex, England, 2007

[5] P. Fan, "Spreading sequence design and theoretical limits for quasisynchronous CDMA systems," *EURASIP J. on Wireless Communications and Networking*, pp. 19 – 31, no. 1, 2004

[6] L. Feng, P. Fan, X. Tang, and K.-K. Loo, "Generalized pairwise Z-complementary codes," *IEEE Signal Processing Letters*, vol. 15, pp. 377-380, 2008

[7] M. J. E. Golay, "Multislit spectroscopy," *J. Opt. Soc. Amer.*, vol. 39, pp. 437 –444, 1949

[8] M. J. E. Golay, "Complementary series," *IRE Trans. Inform. Theory*, vol. 7, pp. 82–87, №2, 1961.

[9] L. Hanzo, M. Munster, B. J. Choi, and T. Keller, *OFDM and MC-CDMA for Broadband Multi-User Communications, WLANs and Broadcasting*, John Wiley & Sons and IEEE Press, New York, NY, USA, 2003

[10] S. P. Maity, M. Mukherjee, "A high capacity CI/MC-CDMA system for reduction in PAPR," in Proc. *Seventh Annual Communication Networks and Services Research Conference*, pp.297-304, Moncton, New Brunswick, Canada, 11-13 May, 2009

[11] A. Milev and V. Mutkov, "Research over correlation properties of families optical orthogonal codes," in Proc. *International Conference on CompSysTech'09*, pp. III.20.1- III.20.6, Ruse, Bulgaria, EU, June 2009

[12] D. Y. Peng and P. Z. Fan, "Generalized Sarvate bounds on periodic autocorrelations and crosscorelations of binary sequences," *Electronics Letters*, vol. 38, no. 24, pp. 1521 – 1523, 2002

[13] T. R. Qureshi, M. D. Zoltowski and R. Calderbank, "A MIMO-OFDM channel estimation scheme utilizing complementary sequences," in Proc. *The 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.2677-2680, Taipei, Taiwan, 19-24 April, 2009

[14] R. G. Tsakov, V. A. Mutkov, N. R. Nikolov and L. A. Staneva, "A survey of the correlation properties of the generalized Barker codes", unpublished

[15] R. G. Tsakov, V. A. Mutkov, N. R. Nikolov and L. A. Staneva, "An improved method for synthesis of families of Costas arrays", unpublished

[16] C.-C. Tseng and C. L. Liu, "Complementary set of sequences," *IEEE Trans. Inf. Theory*, vol. 18, pp. 206–209, Sept. 1972.

[17] M. D. Zoltowski, T. R. Qureshi and R. Calderbank, "Complementary codes based channel estimation for MIMO-OFDM systems," in Proc. *Forty-Sixth Annual Allerton Conference*, pp. 133-138, Allerton House, UIUC, Illinois, USA, 23-26 Sept., 2008

Low Message Overhead Clustering Algorithm for Wireless Sensor Networks

Batista, Othon Marcelo Nunes; Giozza, William Ferreira

Grupo de Redes Ópticas e Wireless (GROW) – Universidade Salvador (UNIFACS)
Rua Ponciano de Oliveira, 126 – Rio Vermelho
CEP: 41.950-275 – Salvador – BA – Brazil
othonb@yahoo.com, william.giozza@gmail.com

***Abstract*—A Wireless Sensor Network (WSN) is defined by a set of sensor nodes and a Base Station (BS). Typically, sensor nodes are distributed over a geographical area, sense some information, process it and send it to the BS via a wireless link. Since sensor nodes are small, may be distributed in large scale or in a dangerous area, their battery are small and may not be recharged or replaced. So the network lifetime is prolonged when the battery energy is used wisely. The function of the BS is to receive information sensed by the nodes, therefore it's a sink. One way to send information is to cluster the WSN, dividing the sensor nodes in two exclusive sets: Cluster Heads (CH) or Cluster Members (CM). Therefore, the clustered WSN presents two types of communication: one from CM to CH, inter-cluster communication, and other from CH to BS, intra-cluster communication. Once a WSN is clustered, energy consumption and coverage are improved. The first step in clustering a WSN is a CH election algorithm. We propose a distributed clustering algorithm for WSN named Low Message Overhead Clustering Algorithm (LMOCA) that has the goal to spend as minimum energy as possible on the clustering process. Beyond that it is very intuitive and easy to implement.**

***Index Terms*— Wireless Sensor Networks, Clustering Algorithm, Simulation Results.**

I. INTRODUCTION

A Wireless Sensor Network (WSN) is defined by a set of sensor nodes and a Base Station (BS). Typically, sensor nodes are distributed over a geographical area, sense some information, process it and send it to the BS via a wireless link. Since sensor nodes are small, may be distributed in large scale and in a dangerous area, their battery are small and may not be recharged by any means. So the network lifetime is prolonged when the battery energy is used wisely. The BS may be a computer with a Universal Serial Bus (USB) dongle to communicate with the sensor nodes. The function of the BS is to receive information sensed by the nodes, therefore it's a sink.

The simplest way to send information to the BS is denominated direct diffusion. As soon as a sensor node senses any information, it is sent directly to the BS in a one hop communication. Each sensor node must configure its transmission power to reach the BS to implement direct diffusion. If a sensor node is too far, much energy is spent and the probability of a collision is bigger. The sensor nodes deaths occur from the farthest to the nearest ones. Direct diffusion forms a plain topology.

Another way to send information to the BS is via multi hop routing. Each sensor must configure its transmission power to reach one or more neighbor sensor node as close as possible. The sensed information is sent to the BS via intermediate sensor nodes, in a multi hop way. The energy spent to send an information to the BS is the summation of the energies spent to send the information to the sensor nodes in a route toward the BS. If a sensor node is very near to the BS, it spends more energy than the others, since it aggregates informations from all other nodes. The sensor nodes deaths occur from the nearest to the farthest ones.

Coverage is the degree of the geographical area in which the sensor nodes were distributed that are in fact being sensed. As the sensor nodes die, coverage is gradually lost. When the information is sent in one hop or multi hop, the coverage is lost really fast, only varying the direction, from the farthest sensor node to the base station or from the nearest sensor node to the base station, respectively.

Another way to send information to the BS is by clustering the WSN. Clustering a WSN means to divide it in two sets of nodes: Cluster Heads (CH) and Cluster Members (CM). This division happens also in the communication level: one communication occurs from CM to CH, intra-cluster communication, and the other from CH to BS, inter-cluster communication. There are several techniques used to cluster a WSN. They may be divided in two groups: localized and distributed approaches. In localized approaches, nodes decide locally, without exchanging messages, if they are CH or CM. In distributed approaches, nodes decide if they are CH or CM by exchanging messages with a distributed algorithm. Once a WSN is clustered, energy consumption and coverage are better than direct diffusion or multi hop routing.

In this sense, we propose a distributed clustering algorithm for WSN named Low Message Overhead Clustering Algorithm (LMOCA). Its goal is to spend as minimum as possible energy in the clustering process by exchanging few messages between sensor nodes. Beyond that, it is very easy to implement, what is good for sensor nodes, given their small computing power and resources.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_96, © Springer Science+Business Media B.V. 2010

II. Related Works

Over twenty clustering schemes were proposed from 2001 to 2005. They may be classified in three types: energy-efficient, k-tree and management schemes. They are compared to Time Division Multiple Access (TDMA) dynamic cluster scheme. This scheme is suitable only sensor nodes distributed in an area forming a grid. Its goals are: to locate CH at the center of a cluster, to elect CH periodically to average power consumption among sensor nodes and not occur duplicated transmission of data through multiple routes [11].

Low Energy Clustering Hierarchy (LEACH) elects CH with a probabilistic localized approach. This is very energy effective, since the radio is not used. It uses TDMA in intra cluster communications and Direct Sequence Spread Spectrum (DSSS) in inter cluster ones. With LEACH, the lifetime of the WSN is divided in turns, each turn is formed by a cluster head election and some intra and inter cluster communication. LEACH spends about eight times less energy than direct diffusion or multi-hop routing communications [12].

There is an improvement of LEACH denominated Energy Efficient Clustering Scheme (EECS) that produces a near uniform distribution of CH and balances the loading among them. EECS showed to prolong network lifetime as much as 35% of LEACH [19].

Clustering is not always the best scheme to make good use of energy in WSN. Clusters must be within isoclusters of monitored phenomenon. An analytical analysis shows that five hops clusters, instead of usual single hop clusters, can provide near to optimal network performance [13].

A comparison of representative clustering techniques and recent developments in connectivity, Medium Access Control (MAC) design, cluster head election periodicity, node duty cycle, optimal cluster size and node synchronization are presented in [9].

Hybrid Energy-Efficient Distributed clustering (HEED) doesn't make any assumption about node distribution, node density or node location. It terminates the clustering process in O(1) iterations and is independent of WSN topology or size. It has a low message exchange and processing cycles and a good cluster head distribution across the WSN [14].

An algorithm for electing cluster heads using maximum residual energy with the goal is to even distribute energy consumption at the overall WSN to obtain the longest lifetime is showed in [15]. The lifetime is expressed as to both the maximum last node dying time and the minimum time difference between the last node dying and the first node dying. Results showed that lifetime can be prolonged when CH are elected with an optimal value of a predefined electing coefficient.

III. Problem Statement

Initially, we present the the assumptions made in LMOCA and state the problem of clustering a WSN.

A. Assumptions

The clustering algorithm was simulated with this assumptions:

- sensor nodes are homogeneous;
- sensor nodes have different transmission power levels, at most 10, but use only one at time;
- frequency used is 2.4 GHz;
- antenna height is 0.03048m;
- receiver sensitivity is -98dBm;
- Friis and Two ray ground are the propagation models used;
- all sensor nodes are initially CM;
- there are neither collisions or loss of packages, since the objective is to test the algorithm;
- BS is located at coordinate (0, 0).

Some of these parameters were gathered from MICA sensor nodes specifications sheet available at [18].

A WSN may be designed as a graph, defined by G = (V, E). V is a finite set of nodes and E is a finite set of edges that connects nodes from V. To apply this definition to WSN, it must be stated that sensor nodes are elements of V and connections based in the sensor node transmission range are the elements of E. A common assumption is that a sensor node transmission range is a Unit Disk Graph (UDG), a circle with the sensor node located at the center and the limits of its transmission range in the circumference. This defines two distinct areas: with connectivity (inside the circle until the circumference) and without connectivity (outside the circle). UDG in wireless networks may be simulated using Friis or two-ray ground propagation models [8].

B. Clustering Problem Statement

Considering a WSN as a graph G may lead us to define clusters by dividing the network in two complementary sets: CH set and CM set. CH set is formed by sensor nodes that are the logical center of clusters. The quantity of elements in the CH set is the quantity of clusters. CH are regular sensor nodes with additional responsibility to aggregate information of a cluster and send it towards BS. In the other hand, elements of CM set, in therms of communication, only send information to its CH.

IV. Cluster Head Election Algorithm

The first step in a clustering protocol for WSN is a CH election algorithm. Since there may be a huge amount of sensor nodes and energy consumption must be minimum, a distributed approach with few message exchanges is desired.

LMOCA has the goal to elect cluster heads with a low message exchange overhead. For each node only three messages are sent to decide if they are CH or CM. At the beginning of the process, node degree information is used to estimate if a sensor node is a CH or a CM. After the cluster heads are elected, a rearrangement of the network is done, by grouping CM to their nearest CH. Hence, CM transmission range is optimized to reach only its CH.

The algorithm is divided in five phases:

phase 1: node degree calculation;

phase 2: node degree exchange;

phase 3: initial clusters;

phase 4: network rearrangement;

phase 5: cluster member transmission range adaptation.

For testing purposes, the phases were configured to run for 5 seconds each. In the future, we intend to find the best time that suits the protocol.

This algorithm was simulated in Castalia, a discrete event WSN [16] simulator based in OMNET++ [17]. For a better explanation, one example is presented in a step-by-step basis with the aid of the WSN graph shown in Fig. 1. This graph represents 100 sensor nodes in an area of 900m^2 (30m x 30m). Sensor nodes are white circles with a number (sensor node identification number). The red circle in coordinate (0, 0) is the BS.

Sensor nodes have a initial transmission power of 0 dBm. This transmission power applied to Friis and Two Ray Ground propagation models lead to a range of approximately 8.59m. This range indicates the distance until the transmission power is reduced to the receiver sensitivity of -98 dBm. Also, it indicates the radius of each cluster.

The five phases describing the operation of LMOCA are presented as follows.

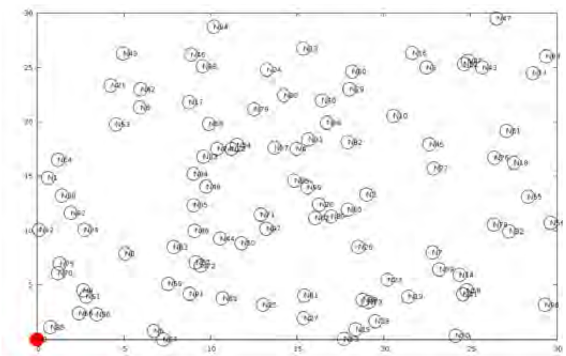


Fig. 1. Example WSN Graph.

A. Node Degree Calculation

The first phase consists of one message sent by every sensor node. This message contains only the sensor node identification number. After sending the message, all sensor nodes keep their receivers on to get messages from the neighbors.

Each message received by a sensor node is counted and the node identification numbers are stored in an array of neighbors. The quantity of messages received is the node degree.

B. Node Degree Propagation

After calculating degree, each sensor node must propagate it to its neighbors. To do so, a sensor node sends a message with the previously calculated degree and the node identification number. Once more, all sensor nodes must keep their receivers on to get the node degree messages.

Received node degrees are stored in an array in a position correspondent to the node identification number in the neighbors array. At the end of this phase, all sensor nodes know their neighbors and neighbour degrees.

C. Initial Clusters

The initial clusters are formed by nodes with higher degree. Since all sensor nodes are initialized as CM, they must choose a node as a CH, if one exists. Each sensor node, checks its node degree array and sends a message to the highest degree sensor node. This message is called DOMINATE ME.

Once a sensor node receives a DOMINATE ME message, it changes its status to CH and stores sensor node identification number in an array of CM sensor nodes.

At the end of this phase if a node neither generate or receive a DOMINATE ME message, it self elects CH. This guarantees that, all nodes are either CM or CH at the end of the clustering phase.

The algorithm could stop after this phase, however, as Fig. 2 shows, the clusters are not well distributed in the geographical area, what will raise the interferences and collisions between clusters. So, there's a needing to reorganize clusters to minimize interference and maximize energy usage. Notice that cluster heads are the red triangles and cluster members are blue circles.

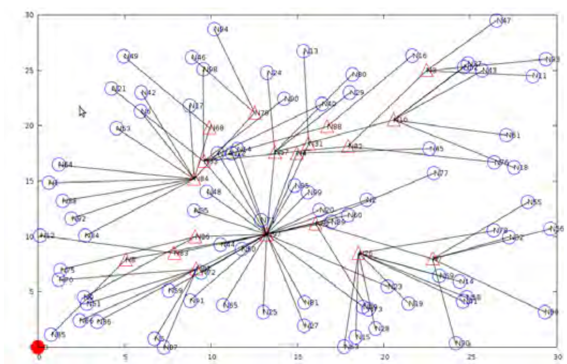


Fig. 2. Initial CDS of the graph presented in Fig. 1.

D. Network Rearrangement

The clustered WSN obtained until now has some clusters formed by members very dispersed in the area. So, cluster heads must send a message that will be received by all sensor nodes in their range. This message will be used by the cluster members to verify, by the received signal strength intensities, which is the nearest cluster head.

E. Cluster Member Transmission Range Adaptation

Once the messages of CH are received, the CM adjust their transmission power to cover the distance to the nearest CH and send a DOMINATE ME message to it. The WSN obtained at the end of this phase contains very well distributed clusters, which must aid in the future inter and intra-cluster communications, regarding that the interferences decrease with the transmission power adjustment (Fig. 3).

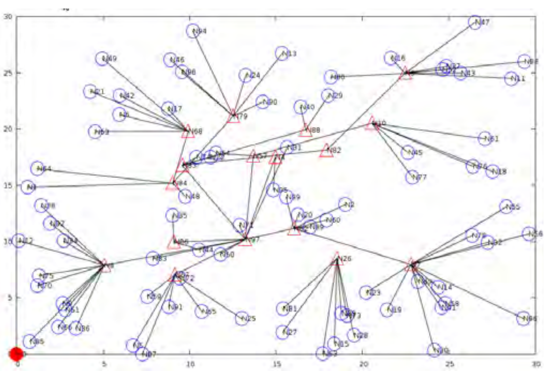


Fig. 3. WSN at the final phase of the proposed algorithm.

V. ALGORITHM ANALYSIS

In this section, we present and prove some lemmas regarding LMOCA.

Lemma 1. LMOCA message complexity is O(1) for each node.

Proof. The quantity of messages sent by each node is, at most, four. CM nodes send three messages, which are:

- an initial message used to calculate node degree;
- a second message to inform calculated node degree;
- a third message to be a member of a cluster (DOMINATE ME);

In the other hand, CH nodes send four messages, which are these three for CM nodes and and additional message to construct well distributed clusters.

Therefore, message complexity for each sensor node is O(1).

Lemma 2. LMOCA message complexity over the network is O(n).

Proof. According to Lemma 1, in the worst case, a sensor node sends four messages until the network is clustered. CM nodes send three messages and CH nodes send four messages. The total quantity of messages sent by all nodes is given by formula 1. In formula 1, $Quantity_{Messages}$ is the total of message for the network, $Quantity_{CH}$ is the number of CH and $Quantity_{CM}$ is the quantity of CM.

Formula 1. Quantity of messages sent to cluster a wireless sensor network.

$$Quantity_{Messages} = 4 * Quantity_{CH} + 3 * Quantity_{CM}$$

Hence, message complexity for the whole network in the worst case is O(n). n is the total number of nodes in the network.

Lemma 3. At the end of LMOCA, a sensor node is either a cluster head or a cluster member.

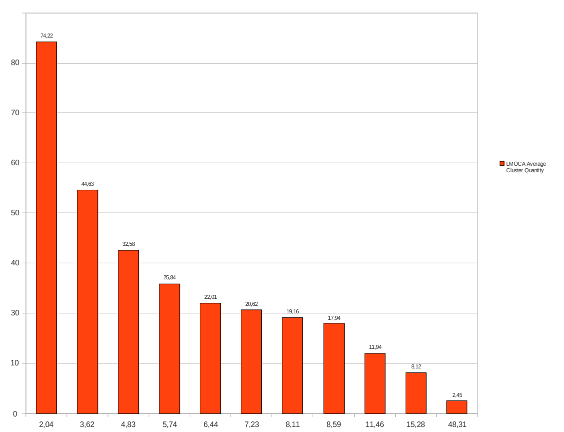
Proof. At the end of the clustering phase, if a sensor node neither generate or receive a DOMINATE ME message, it self elects a CH. This guarantees that, all nodes are either CM or CH at the end of the clustering phase.

So, at the end of the execution of LMOCA, all nodes are either CH or CM.

VI. SIMULATION RESULTS

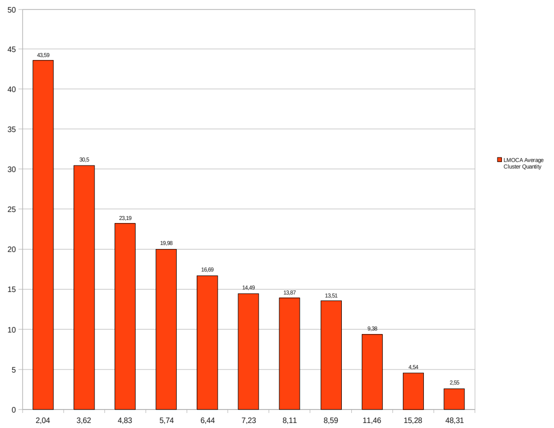
LMOCA was executed 100 times in five different scenarios. Simulations were executed with sensor nodes located in different positions. For each scenario, the quantity of CH was observed to be related to the transmission power. Parameters used in simulations were the same presented in section III of this paper. The only differences are one additional power level and the quantity of sensor nodes per scenarios: 100, 50, 10 and 2, respectively. Graphs 1, 2, 3 and 4 have in axis X the cluster radius, which is proportional to transmission power applied to Friis and Two Ray Ground propagation models, and in axis Y the quantity of CH obtained by LMOCA.

The results of the first simulated scenario are presented in Graph 1. In this scenario, 100 sensor nodes were used, each simulation in a aleatory different position. It shows that the quantity of CH is inversely proportional to transmission power. For instance, if we increase the transmission power to the maximum, which leads to a cluster radius of 48,31m, we decrease the quantity of CH to approximately 2.



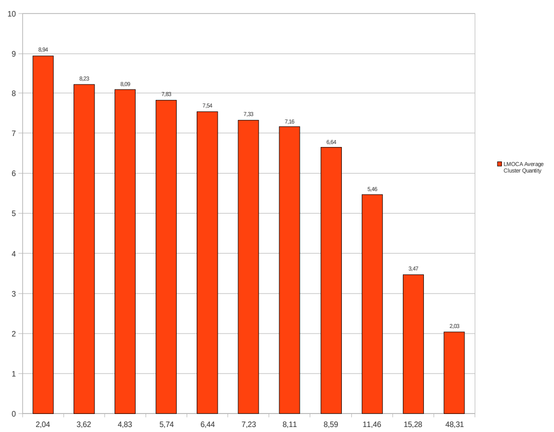
Graph 1. Quantity of CH after 100 simulations for 100 sensor nodes.

In Graph 2, results obtained in second simulation are presented. Scenario for this graph is composed with 50 sensor nodes in different positions for each simulation. The relation of the quantity of CH to be inversely proportional to transmission power still holds.



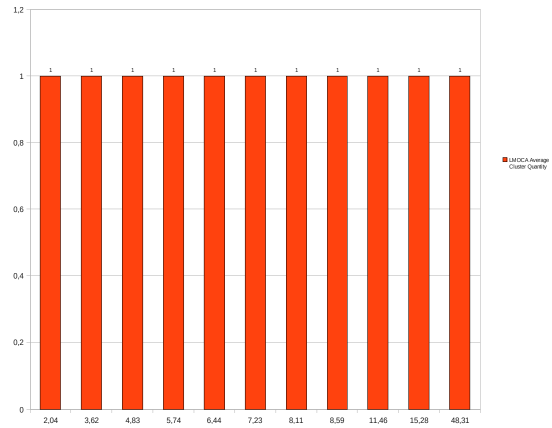
Graph 2. Quantity of CH after 100 simulations for 50 sensor nodes.

The results of the third simulated scenario are presented in Graph 3. In this scenario, 10 sensor nodes were used, each simulation in a aleatory different position. The quantity of CH is inversely proportional to transmission power.



Graph 3. Quantity of CH after 100 simulations for 10 sensor nodes.

The fourth and last scenario analyzed has only 2 sensor nodes. This is to show that LMOCA runs correctly even with the minimum quantity of sensor nodes. Only one sensor node is CH. The graph 4 shows the results for this situation.



Graph 4. Quantity of CH after 100 simulations for 2 sensor nodes.

Although not graphically shown in this paper, the scenario with only one sensor node were simulated. All the times the quantity of CH were 1. This also shows that LMOCA works for this special situation.

VII. Conclusion

This paper presented a proposal for an algorithm to elect cluster heads in a WSN. For each sensor node, at most, four messages are sent to form well distributed clusters, which minimizes the interferences, collisions and, consequently, saves energy. Some lemmas were proved for the LMOCA and simulation results were presented. In these results, we observed that the quantity of CH depends on the transmission power. Bigger transmission powers lead to fewer CH. LMOCA is part of a clustering protocol and by this way, future works consist of finding a CH rotation policy based in residual energy and the integration of an intra-cluster and inter-cluster communication scheme to this algorithm.

References

[1] Brooks, R. R., Ramanathan, P. and Sayeed, A. M. (2003) “Distributed Target Classification and Tracking in Sensor Networks” Proceedings of the IEEE, Vol. 91, NO. 8.

[2] Werner-Allen, G., Johnson, J., Ruiz, M., Lees, J. and Welsh, M. (2006) “Monitoring Volcanic Eruptions with a Wireless Sensor Network”, Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2006). Seattle, USA.

[3] Liu, T., Sadler, C. M., Zhang, P. and Martonosi, M. (2004) “Implementing Software on Resource-Constrained Mobile Sensors: Experiences with Impala and ZebraNet”, Proceedings of the The Second International Conference on Mobile Systems, Applications, and Services (MobiSYS'04). Boston, Massachusetts, USA.

[4] Krishnamurthy, L., Adler, R., Buonadonna, P., Chhabra, J., Flanigan, M., Kushalnagar, N., Nachman, L. and Yarvis, M. (2005) "Design and Deployment of Industrial Sensor Networks: Experiences from a Semiconductor Plant and the North Sea", Proceedings of the 3rd ACM Conference on Embedded Networked Sensor Systems (SenSys). San Diego, California, USA.

[5] Szewczyk, R., Mainwaring, A., Polastre, J., Anderson, J. and Culler, D. (2004) "An Analysis of a Large Scale Habitat Monitoring Application", Proceedings of the 2rd ACM Conference on Embedded Networked Sensor Systems (SenSys). Maryland, USA.

[6] Su, H. and Zhang, X. (2006) "Optimal Transmission Range for Cluster-Based Wireless Sensor Networks with Mixed Communication Modes", Proceedings of the 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'06).

[7] Heinzelman, W. B. and Chandrakasan, A. P. (2002) "An Application-Specific Protocol Architecture for Wireless Microsensor Networks", IEEE Transactions on Wireless Communications, volume I, Issue 4, Oct 2002, p.p. 660 - 670.

[8] Kuhn, F., Moscibroda, T. and Wattenhofer, R. (2004) "Unit Disk Graph Approximation", Proceedings of the 2004 Joint Workshop on Foundations of Mobile Computing, Philadelphia, PA, USA, p.p. 17 - 23.

[9] Younis, O., Krunz, M. and Ramasubramanian, S. (2006) "Node Clustering in Wireless Sensor Networks: Recent Developments and Deployment Challenges", IEEE Network. May/June.

[10] Funke, S., Kesselman, A., Meyer, U. and Segal, M. (2006) "A Simple Improved Distributed Algorithm for Minimum CDS in Unit Disk Graphs", ACM Transactions on Sensor Networks, Vol. 2, No. 3, August, p.p. 444-453.

[11] Dali, W. and Chan, A. C. (2005) "A Survey on Cluster Schemes in Ad Hoc Wireless Networks", Proceedings of the IEEE Mobility Conference. Guang Zhou, China.

[12] Heinzelman, W. R., Chandrakasan, A. and Balakrishnan, H. (2000) "Energy-Efficient Communication Protocol for Wireless Sensor Networks". Proceedings of the 33rd Hawaii International Conference on Systems Sciences .

[13] Vlajic, N. and Xia, D. (2006) "Wireless Sensor Networks: To Cluster or Not to Cluster?", Proceedings of the 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'06).

[14] Younis, O. and Fahmy, S. (2004) "Distributed Clustering in Ad-hoc Sensor Networks: A Hybrid, Energy-Efficient Approach", Proceedings of the. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 1, p. 640.

[15] Wang, W. and Jantsch. A (2006) "A new protocol for electing cluster head based on maximum residual energy". In Proceedings of the IEEE International Cross-Layer Designs and Protocols Symposium.

[16] Pham, H. N, Pediaditakis, D. and Boulis, A. (2007) "From Simulation to Real Deployments and Back". in Proceedings of t2pWSN.

[17] Omnet++ (2009) "OMNET++ Discrete Event Simulation System". http://www.omnetpp.org.

[18] XBow (2009) "Crossbow Technology". http://www.xbow.com. San Jose, CA. USA.

[19] Ye, M., Li, C., Chen, G. and Wu, J. (2005) "EECS: An Energy Efficient Clustering Scheme in Wireless Sensor Networks". the 24th IEEE International Performance, Computing, and Communications Conference. p.p. 535-540.

On the Design of SOHO Networks

Lawan Ahmed Mohammed

King Fahd University of Petroleum and Minerals
HBCC, 31991 Hafr Batin, Saudi Arabia
gumel@hbcc.edu.sa

Abstract — **Small networks such as small office/ home office (SOHO) face the same threats as large enterprise networks. However, they also contend with the challenge of limited resources and budgets for IT expenditures. In such networks environment, the role of administering the system often falls on the business owner or on the default setup when the system was installed. In most cases, the owners do not usually have the time, resources or expertise to work on complex security problems. Similarly, most of the default setup use to have some loopholes which were not initially identified. This paper describes the primary security challenges facing SOHO networks today, and suggests simple easy to use security solutions to resolve these challenges.**

Keywords— **Network security, Design of SOHO, Flooding attack, Security measures.**

I. INTRODUCTION

The development of cost-effective networking technologies has provided large amount of new possibilities for the users in small office and home (SOHO) environments. Some of the benefits of these networks include: Improved efficiency and access to mobile employees, Offers mobility and flexibility, Increase productivity and reduced costs, Optimizing business processes, Increase competitive advantage, Improve customer satisfaction and retention etc. As with all types of networking environments, SOHO environments and technologies are also vulnerable. Security issues such as authentication, authorization, and access control requirements should be considered when designing services for such networks. As SOHO networks can be either wired or wireless, peculiar vulnerabilities associated with the chosen technology should also be taken into consideration. In view of these, implementing and supporting scalable SOHO faces many challenges. For instance, these environments consist of variable hardware and software components, coming from different vendors or manufacturers some being more trustworthy than others. In this case issues to be managed may include integration management, rapid technological expansion and changes etc.

Lightweight SOHO computing devices like Personal Digital Assistants (PDAs), mobile phones, and laptops are also prone to large set of software attacks like the denial of service (DOS) or buffer overflow attacks. These devices when connected to infrastructure based networks like office LAN, make them vulnerable to the attacks similar to that on infrastructure based networks. Addition of mobile, handheld, and portable computing nodes into existing fixed infrastructure based networks increases the heterogeneity of the resulting network, making the problem of intrusion detection even harder especially as these devices can be connected to each other and also to Internet by using wired or wireless technology. For example, wireless technology like Bluetooth, 802.11, IrDA can provide LAN and Internet access to a handheld computer wirelessly through access points. The mobility aspect of these devices makes it harder to protect them. Further, different components and networking technologies support different security mechanisms, which may not be compatible and which have different security characteristics. This implies the need for multi vendor's device and security management. For instance, usually, Access Point (APs) from manufacturers comes with a set of default configuration parameters. These default parameters leave some loopholes for attacks. For example (depending on the manufacturer), most APs have a default administrator password, SSID, channels, authentication/encryption settings, SNMP read/write community strings, etc. Since these default values are available in user manuals, vendor's website and installation guide, they are well known to the general public and may be used by hackers to compromise WLAN security. Furthermore, these environments involves different category of users with different access right and privacy requirements. For example, home users may include parents, children, guests, and external administrators while offices may have employees with different roles. Finally, as the system may involve different type of users, individual skills and expertise challenges should also be considered.

As SOHO is generally small and require fewer infrastructures, attack on this network can easily make the entire system inoperable. In view of these, the focus of the paper is to examine the security threats and requirements especially in the design of wireless SOHO networks. First, the paper provides an overview of SOHO design. Later, it discusses security threats and attacks for these environments. Finally, it presents some methods of preventing the attacks.

II. OVERVIEW OF SOHO DESIGN

Wired SOHO uses either dialup or Ethernet. This consists of network adapters, hub, and simple configuration software. Unlike in wired SOHO, wireless SOHO uses WLAN technology based on the IEEE 802.11 standard. Users can freely move inside or outside the building and still remain connected to the network and other hardware devices on the network. This network components can be set up anywhere in

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_97,

the building. Further, wireless SOHO is very flexible and also allows users to move computers and other devices without the need to reconfigure the network. Flexibility comes in as users can get connected through simple steps of authentication without the hassle of running cables. Also compared to the wired network, wireless network installation costs are minimal as the number of interface hardware is minimal. As illustrated in table 1 below, WLANs are less expensive to install even for relatively modest workforces. Those savings grow along with the size of the workforce as the cost associated with WLAN controllers is further amortized. Further, wireless SOHO does not require any computer running Windows server 2003, it fully supported in Windows XP (even home edition), it is relatively easy for even a novice user to deploy, it does not require the purchase of a server certificate, and also provides simple methods to control or manage workgroup members. These reasons make wireless SOHO more popular than the wired SOHO.

TABLE 1
COST OF WIRED VS WLAN (Source: www.Sonicwall.com)

Equipments	Ethernet	WLAN
Ethernet/PoE Port	$35	$35
Installation	$200	$250
AP	-	$1,300 (50 users)
Controller	-	$15,000 (50 APs)
Installed costs - 100 users	$23,500	$16,585
Installed costs - 1,000 users	$235,000	$46,700

In a typical WLAN configuration, a transmitter/receiver (transceiver) or AP, connects to the wired network from a fixed location using standard Ethernet cable. At a minimum, the Access Point receives, buffers, and transmits data between the WLAN and the wired network infrastructure. A single Access Point can support a small group of users and can function within a range of less than one hundred to several hundred feet. Fig. 1 below shows the basic components of a SOHO wireless network infrastructure.

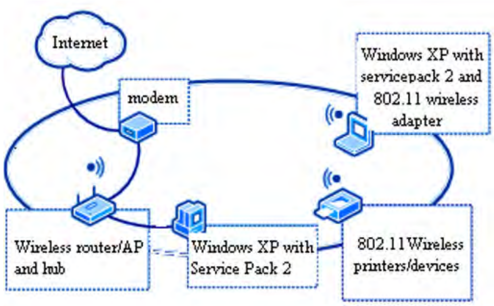


Fig 1 Basic Wireless SOHO Infrastructure

Some of the well-known SOHO benefits, including:

- Improved efficiency and access to mobile employees
- Offers mobility and flexibilty
- Increase productivity and reduced costs
- Optimizing business processes
- Increase competitive advantage
- Improve customer satisfaction and retention. etc

On the negative side, some of the key challenges can be summarized as follows:

- Integration management
- Multi vendors device and security management
- Rapid technological expansion and changes
- Skill and expertise challenges
- Future proofing challenges etc

III. SECURITY THREATS

SOHO computing environment share all of the security issues of traditional networked applications. These include authentication of devices and users, privacy of data or information, defence against malicious code such as viruses, worms, Trojan horses etc, and access control mechanisms. However, the SOHO computing environment combines both wired and wireless network security challenges as the system can be based on both types of environment. When wireless network is employed, information that is usually confined behind a corporate firewall will be winging its way through the air, possibly spending some time on hosted servers or wireless gateways. Further, the techniques of hacking wireless and mobile devices such as laptops, cell phones, PDAs etc is already spreading. In view of these, adding security to such environment presents challenges at different levels. For instance, having a central authority for a single building or even a group of rooms is infeasible because every possible access right will have to be specified for every user. Authenticating the identity certificate of a previously unknown user doesn't provide any access control information. Simple authentication and access control are only effective if the system knows in advance which users are going to access a particular subject or stored information and what their access rights are. Portable handheld and embedded devices have severely limited processing power, memory capacities, software support, and bandwidth characteristics. Also, hardware and software environments are becoming increasingly heterogeneous [1]. In view of these, the security requirements can be roughly divided into three main categories; network security requirements, device security requirements, and finally the user security requirements. However, in this paper we focus more on the network security requirements.

SOHO computing aims to provide users with computer-supported capabilities like traditional network environment. In order to achieve this, correct identification and verification of users is paramount. Authentication is equally important when allowing a user access to a secure file space. Its purpose is to ensure that each entity is correctly matched to its corresponding privileges. If someone is not accurately identified then their identification is left unverified. In wireless SOHO computing environment, the risk of someone fraudulently performing input using authorized user's login details is greater and, for reasons such as this, access to these workspaces and the computers within them may require more physical and information security measures than the normal wired network environment. Details of WLAN vulnerabilities can be found in [2], [3] and [4]. Moreover, one of the main difficulties in designing a secure SOHO environment is in ensuring that the functionality of the environment is not over constrained by security. Brief descriptions of some specific attacks targeting the entire network and its infrastructure when the devices are connected to the SOHO network are given below:

Web Attacks – The web browser is a complex application that can be extended through entire programming languages such as Javascript, Java applets, Flash applications and ActiveX controls. While these technologies make for a richer web experience for the end user, each of them is a new attack vector in a program running inside a firewall. Filtering network traffic cannot protect against all attacks. Further, even Web sites that are legitimate for use in a business context can serve as a source of threats.

Impersonation attacks - An intruder assumes the identity and privileges of another node in order to consume resources or to disturb normal network operation. An attacker node achieves impersonation by misrepresenting its identity. This can be done for example by changing its own MAC address to that of some other legitimate node. Strong authentication procedures can be used to stop attacks by impersonation.

Eavesdropping – This implies the interception of information/data being transmitted over the network. In case of wireless SOHO, when the wireless link is not encrypted, an attacker can eavesdrop the communication even from some few miles away. The attacker can gain two types of information from this attack. The attacker can read the data transmitted in the session and can also gather information indirectly by examining the packets in the session, specifically their source, destination, size, number, and time of transmission. Eavesdropping can also be active; in this case the attacker actively injects messages into the communication medium in order to assist him/her in determining the contents of messages.

Poor network administration - A surprising number of machines are configured with an empty or easy to guess root/administrator password. One of the first things an attacker will do on a network is to scan all machines for empty or commonly used passwords.

System Intrusion – Also known as “Privilege Escalation”, this type of hacking assumes the attacker already has a low-privilege user account on the system. If the system doesn't have the latest security patches, there is a good chance the attacker will be able to use a known exploit in order to gain additional administrative privileges.

TCP/IP protocol flaws - The TCP/IP protocol was designed before we had much experience with the wide scale hacking we see today. As a result, there are a number of design flaws that lead to possible security problems. Some examples include smurf attacks, ICMP Unreachable disconnects, IP spoofing, and SYN floods.

Traffic Analysis – Traffic analysis allows the attacker to obtain three forms of information. The attack primarily identifies that there is activity on the network. Similar to standard radio communications, a significant increase in the amount of network activity serves as an indicator for the occurrence of a large event. Secondly, the attacker can find information about the location of APs in the surrounding area. This is because unless turned off, APs broadcast their service set identifiers (SSIDs) for identification. Thirdly, the attacker may learn the type of protocols being used in the transmission. This knowledge can be obtained based on the size, type and number of packets in transmission over a period of time. Analysis of TCP three-way handshake is described in [5].

Unauthorized Access – this attack involves gaining free access to the network and also using the AP to bypass the firewall and access the internal network. Once an attacker has access to the network, he/she can then launch additional attacks or just enjoy free network use. Although free network use may not be a significant threat to many networks, however, access is a key step in ARP based man-in-the-middle attacks.

AP Overloaded - This attack is based on the observation that in 802.11 a client must be successfully authenticated and associated to an AP before using wireless communication services. AP maintains the client state information in a client-association table. When the table reaches the permitted level of associated clients, the AP start rejecting new association requests. This attack is facilitated by the open system authentication method in 802.11where an AP authenticates anyone who requests authentication. An attack can be launched if the adversary does a large number of associations with an AP, using random MAC addresses. Since an AP can maintain a limited number of associations, this will prevent other stations from joining the AP.

Layer-Level Attack - Every year, network attacks become more widespread, more intelligent and more difficult to detect. Because many of today's attacks are blended attacks which use multiple techniques at different layers to try to infiltrate the network, they can bypass outdated firewalls that only provide packet inspection for network traffic. With packet inspection, firewalls mainly ensure that connections are valid. However, they perform absolutely no inspection beyond the first few networking layers and thus do not concern themselves with the content carried within these data streams, allowing both desired and malicious data through. Further, as the amount of traffic being scanned by firewall as well as the increasing amount of threats and malware lurking in that traffic is quickly becoming more than many firewalls can handle. Inspecting every byte of every packet can overwhelm some firewalls and bottleneck network performance. This not only keeps users from getting the most out of the available bandwidth, but also degrades and disrupts streaming applications like Voice over IP (VoIP) phone and video.

Modern Malware - The goal of malware is money, not recognition, and stealth is highly prized. The methods and techniques for making malware pay expand and evolve at a rapid pace, with the more general risks of infected systems being absorbed into botnets for use in spam campaigns and DDoS attacks, draining bandwidth and resources, and halt device performance, supplemented by the more personal dangers of data theft. A more dangerous malicious program can transmit itself across the network, bypassing some of the network security systems, and potentially damaging other components of the network.

IV. ATTACK SCENARIO

In this section, we demonstrate a simple scenario to show how easily a network with limited resources such as SOHO can be disrupted by ping flooding attack. First, we briefly explain the purpose of ICMP and then show how the vulnerability associated with this protocol and leads to flooding attack.

Internet Control Protocol or ICMP is used to report the delivery of Internet Protocol (IP) echo packets within an IP network. It can be used for network trouble shooting purposes to show when a particular end station is not responding, when an IP network is not reachable, when a node is overloaded or when an error occurs in the IP header information etc [6].

PING flood attack: One of the features of TCP/IP is fragmentation; it allows a single IP packet to be broken down into smaller segments. Attackers take advantage of that feature when they found that a packet broken down into fragments could add up to more than the allowed 65,536 bytes. Many operating systems cannot handle oversized received packet, so they froze, crashed, or rebooted.. We demonstrated this as shown in Fig. 2, 3, and 4 below.

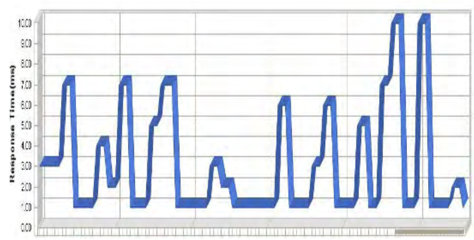


Fig. 2 The response time at the beginning of the flood attack

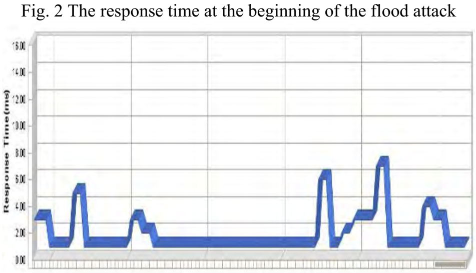


Fig. 3 The response time at the middle of the flood attack

The excess and unwanted ICMP packets that flood the target computer buffers have caused this lack of response as shown in fig 5below.

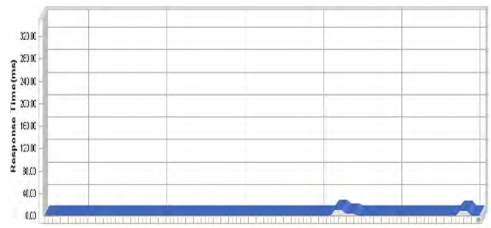


Fig. 4 The response time at the end of the flood attack

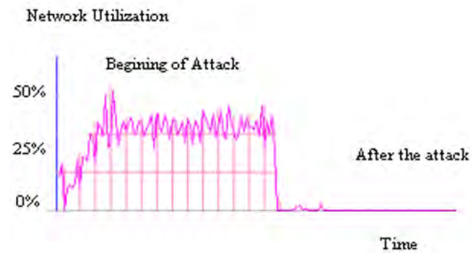


Fig. 5. The status of a victim before and after the flood attack.

Some types of attacks that are associated with ICMP are documented in [7], [8], and [9]:

ICMP DOS Attack: This is a denial-of-service attack in the sense that the attacker could use either the ICMP “Time exceeded” or “Destination unreachable” messages to force a host to immediately drop a connection. An attacker can make use of this by simply forging one of these ICMP messages, and sending it to one or both of the communicating hosts. Their connection will then be broken. The ICMP "Redirect" message is commonly used by gateways when a host has mistakenly assumed the destination is not on the local network. If an attacker forges an ICMP "Redirect" message, it can cause another host to send packets for certain connections through the attacker's host.

ICMP packet magnification (or ICMP Smurf): *In Smurf attack,* which is one of the most effective in the category of network-level attacks against hosts, the attacker sends a large amount of ICMP echo (ping) traffic at IP broadcast addresses, all of it having a spoofed source address of a victim. If the routing device delivering traffic to those broadcast addresses performs the IP broadcast, most hosts on that IP network will take the ICMP echo request and reply to it with an echo reply each, multiplying the traffic by the number of hosts responding. On a multi-access broadcast network, there could potentially be hundreds of machines to reply to each packet. The “smurf” attack's close relative is called “fraggle”, which uses UDP echo packets in the same way as the ICMP echo packets; it was a simple re-write of “smurf”.

Inverse Mapping - Inverse Mapping is a technique used to map internal networks or hosts that are protected by a filtering device. In this attack an attacker sends an ICMP reply message to a range of IP addresses presumably behind a filtering device. Upon receiving the series of ICMP reply messages, since the filtering device does not keep state of the list of ICMP requests, it will allow these packets to their destination. If there is an internal router, the router will respond with a ICMP "Host Unreachable" for every host that it cannot reach, thus giving the attacker knowledge of all hosts which are present behind the filtering device.

Crafted ICMP Messages: An attacker can send a specially crafted ICMP message to an affected system. An attacker who successfully exploited this vulnerability could cause the affected system to reset existing TCP connections, reduce the throughput in existing TCP connections, or consume large amounts of CPU and memory resources. These attacks, which only affect sessions terminating or originating on a device itself, can be of three types: 1. Attacks that use ICMP "hard" error messages, 2. Attacks that use ICMP "fragmentation needed and Don't Fragment (DF) bit set" messages, also known as Path Maximum Transmission Unit Discovery (PMTUD) attacks, 3. Attacks that use ICMP "source quench" messages. Successful attacks may cause connection resets or reduction of throughput in existing connections, depending on the attack type.

V. DEFENSE MECHANISMS

The following countermeasures can be taken to minimize or thwart the attacks mentioned above.

Against Unwanted Web Contents - An effective way that can be implemented to block Web-based threats is content filtering designed to block unwanted file types. Blocking file types based on their content can be useful in preventing some types of Web threats from entering a network, particularly files that are traditionally known to be associated with malware, such as .scr or .pif. These systems can also block file types that are generally not used in a legitimate business context, such as .mp3, .jpg or .mov files.

Against Access to Harmful Websites - Another option that should be considered is the deployment of URL filtering tools that will block access to non-approved Web sites. Many organizations have deployed these filters, albeit with varying levels of success. While URL filters can be useful, they can rarely keep up with the new threats that enter the Web on an hourly basis and for which no signature has been created in the tool. Further, URL filters can generate significant levels of false positives – blocking Web sites that appear to be suspicious but might have a legitimate business purpose.

Against Malware - A new approach is required to thwart the escalating threats posed by targeted malware and botnets that render traditional security techniques obsolete. Deploying an anti-botnet protection system that uses global intelligence combined with local analysis and control ensures that botnets do not promulgate within the enterprise network. An adequate protection system scans all network traffic and executes relevant traffic streams in virtual victim machines – safe from real clients – and conclusively determines if botnet malware is present. This is the only real defense against network-borne attempts at spamming, spear-phishing, coordinated attacks, identity theft, or theft of intellectual property. It eliminates the underlying cause for all these events—botnets

Against Spoofing - Attackers launching spoofing usually hide the identity of machines they used to carry out an attack by falsifying the source address of the network communication. This makes it more difficult to identify the sources of attack traffic. It is therefore important to use network switches that have MAC binding features that store the first MAC address that appears on a port and do not allow this mapping to be altered without authentication. To prevent *IP spoofing*, disable source routing on all internal routers and use ingress filtering. *DNS spoofing* can be prevented by securing the DNS servers and by implementing anti-IP address spoofing measures. Some vendors have added access control lists (ACL), implemented through MAC address filtering, to increase security. MAC address filtering amounts to allowing predetermined clients with specific MAC addresses to authenticate and associate. While the addition of MAC addresses filtering increases security, it is not a perfect solution given that MAC addresses can be spoofed. Also, the process of manually maintaining a list of all MAC addresses can be time consuming and error prone. Therefore MAC address filtering is probably best left for only small and fairly static networks [10].

Against Flooding Attack - A combination of Host-based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System (NIDS) can greatly help against this attack. HIDS can be placed on critical servers and NIDS can be placed on one or more network segments. Signature detection scheme would be good at detecting any known attacks. Alerts arising from any suspicious activity can be intimated to the administrator immediately. NIDS reactions can also be TCP resets, IP session logging and Blocking. HIDS approach looks into log files, file checksums and intercepting requests to the operating system for system resources before they can proceed. Signatures and generic rules help in anomaly detection. Open server ports can also be monitored for excess or abnormal traffic. Firewalls are an excellent form of protection; however, they must leave some ports open to allow the operation of the web, mail, ftp, and other Internet based services, and which are the paths exploited by most of the vulnerabilities. A more effective measure may involve a hybrid IDS system. This is a distributed system where each and every node in the network collectively cooperated to form a team of IDS agents for the entire network. This can easily be done by placing an IDS intelligent agent on every node in the network. Each IDS agent runs independently and monitors local activities (user, system, or communication) in the network. It will then program to detect any intrusion from local traces (especially communication activities). If any anomaly or suspected activity is detected, a report is send to the host server and other nodes. These individual IDS agents form the system's IDS to defend the entire network.

Against Layer-Level Attack - One of the simple was to secure SOHO network today is the use of Unified Threat Management (UTM) tools that layer on antimalware protection, content filtering, antispam and intrusion prevention, because deploying a single, multi-function device reduces costs and simplifies configuration. Thus, UTM is a network security tool providing a new level of scrutiny to network traffic passing into an organization. Simply put, UTM firewalls combine the effectiveness of various point defences to add protection at every networking layer. The power of UTM comes from its simplicity: a single appliance that provides a powerful defence against a wide range of security threats, while extending beyond, the capabilities of regular packet inspection firewalls. This makes network protection more complete, affordable and easy to manage. For optimal performance while maintaining maximum security, solutions such as reassembly-free deep packet inspection UTM help to deliver throughput that won't bog down performance. Clever algorithms help to reduce latencies to non-noticeable levels. Also, advances in microprocessor technology, especially those designed for network processing, allow UTM appliances designed for SOHO environment to gain incredible network efficiency.

Alternatively, a firewall can be considered when connecting that network to a second network (such as virtual private network or VPN) which has a less stringent security policy. Implementing firewall along with VPN over the SOHO LAN/WLAN can strengthen the system's security. IPSec can be used as the security protocol and the secure tunneling with authentication and encryption. In [11], we demonstrated this idea using a simple network topology. Two nodes A and B had to access database and web pages on a Server through a network with firewall and VPN server implemented. The firewall would not permit database access, but would allow HTTP or web service access. A compulsory VPN tunnel is implemented between the routers (source router and destination routers) that connect node A to the Server. So even if the firewall blocks a service, a node with VPN tunnel enabled can go through the firewall as data communication happens as encapsulated packets (packet within a packet).

Against Wireless Breaches -When wireless network is deployed, it is recommended to check all default settings, including administration password and access path, and change them or disable functions as needed. Turn off all access point services that are not being used (ftp, http, etc…). Also WEP encryption must be turned on and must use 128 bit keys. Moreover, SSIDs must be changed from defaults, must not reflect any information about the organization or location of the Access Point, and SSID cloaking must be turned on. MAC address filtering must be used to restrict wireless access to only approved hardware addresses. The AP locations should be in interior positions and away from exterior walls and windows, ensure they are placed in secured locations to prevent unauthorized physical access.

Disaster Recover Plan - No matter how secure the network is, the need for a disaster recovery solution is essential. Major disasters have demonstrated how exposed small businesses can be to unexpected events. Building fires, broken water pipes, power outages, equipment failures, terrorist, or even lost or stolen handheld devices, can mean disaster for small businesses networks such as SOHO. Modern UTM appliances can feature integrated IPSec or SSL VPN capabilities. SSL VPNs are best suited for secure remote access during an emergency because they allow workers and partners to connect safely to corporate network resources using a Web portal, without having to pre-install clients. Modern continuous data protection (CDP) solutions can automatically backup data and applications to discs, avoiding the complexity and human error involved with tape backup. Backup to a secure secondary business location or third-party site means business systems can be restored and operational even in the primary site is compromised. Bare metal recovery (BMR) technology enables entire operating systems, such as database or file servers, be recovered to new or different hardware platforms if the original device can't be restored.

VI. CONCLUSION

This paper described some vulnerabilities associated with SOHO networks. It shows that implementation of wireless SOHO posses more loopholes than wired SOHO if proper security measures are not taken into consideration. The paper discuses some essential areas required for a complete and robust protection policy. These include anti-malware protection on all platforms, along with firewalling, intrusion prevention, network access control, data leak prevention and data encryption among others.

ACKNOWLEDGMENT

The author wishes to acknowledge Hafr Al-Batin Community College and King Fahd University of Petroleum and Minerals (KFUPM) Saudi Arabia for their support in providing the various facilities utilized in the presentation of this paper.

REFERENCES

[1] L. Kagal, F. Tim, and J. Anupam, "Trust-Based Security in Pervasive Computing Environments", *Computer* v34, n12, Dec 2001, pp:154-157T.

[2] T. Takahashi, "WPA passive dictionary attack overview" [Online]: http://www.3c.com.vn/Uploaded/huyenctt/New%20Folder/WPA_Passive_Dictionary_Attack_Overview.pdf

[3] P. Liza, "WPA PSK Crackers: Loose Lips Sink Ships", [Online]. www.wi-fiplanet.com/tutorials/article.php/3667586

[4] E. Tews 'Gone in 900 Seconds, Some Crypto Issues with WPA', *PacSec 2008 Conference*, Nov. 2008, Tokyo, Japan.

[5] W. Donald, "Wireless Security Threat Taxonomy", I*EEE Workshop on Information Assurance*. US Military Academic West Point N.Y, June 2003, pp: 76-83.

[6] Internet Control Message Protocol (ICMP)., [Online]: http://www.erg.abdn.ac.uk/users/gorry/course/inet-pages/ icmp.html

[7] ICMP Attack, [Online]: http://www.javvin.com/networksecurity/ ICMPAttacks.html

[8] Cisco White Paper: "Crafted ICMP Messages Can Cause Denial of Service " [Online]: http://www.cisco.com/warp/public/707/cisco-sa-20050412-icmp.shtml

[9] SANS White Paper, "ICMP Attacks Illustrated" [Online]: http://www.sans.org/info/36914

[10] L. A. Mohammed and B. Issac, "Detailed DoS Attacks in Wireless Networks and Countermeasures" *International Journal of Ad Hoc and Ubiquitous Computing, (IJAHUC*), Vol 2. No 3, pp. 157-166, 2007

[11] B. Issac and L. A. Mohammed , "War Driving and WLAN Security Issues Attacks, Security Design and Remedies", *Information Systems Management*, Vol 24, pp. 289-298, 2007

Study of UMTS Handover Scenarios

B. Issac
Swinburne University of Technology (Sarawak Campus), Malaysia
bissac@swinburne.edu.my

***Abstract*— Universal Mobile Telephone system (UMTS) is a third generation mobile and wireless communications standard which is also being developed into a fourth generation technology. We wanted to study the performance of such a wireless cellular network during soft and hard handover scenarios with single and multiple towers. An attempt is thus made to use existing simulation models to study the related performance graphs as the mobile node moves around.**

I. INTRODUCTION

The European countries have chosen Universal Mobile Telephone system (UMTS) as one of the standards for the third generation mobile communication system. The third generation in the Nordic countries has been preceded by the first generation mobile telephone system NMT (Nordic Mobile Telephone), and the second generation GSM (Global system for Mobile Communication). UMTS is planned to coexist or work with 2G for several years, as NMT coexisted with GSM from 1992 to 2002, when NMT finally was phased out. As with NMT and GSM, GSM handsets will not be compatible with the UMTS. On the other hand, the different networks will not be totally separated as UMTS handsets will be compatible with GSM providing a dual mode of communication. This is because UMTS, as NMT, will provide less coverage than GSM, probably only rural areas with high density of users.

UMTS is defined by European Commission [1] as: "... a third-generation mobile and wireless communications system capable of supporting in particular innovative multimedia services, beyond the capability of second generation systems such as GSM ..." . This definition is not that specific on what new services can be expected by the users. International Telecommunication Union is slightly more specific as they define UMTS by higher capacity and enhanced network functionalities that would allow more demanding services and applications, like multimedia. For an end-user, the major difference between 3G and 2G is underlined by these definitions to be the increased user value created by a broad service offering. UMTS provides key features for end-users provided as – Options or facilities for Internet and other multimedia-rich applications, a good range of content services and global roaming capability.

UMTS will provide basic services as with GSM and they are – voice, Short Messaging Service (SMS), roaming and billing. So multimedia services would be developed by operators as well as service providers and that will create added interest and it can be expected that different forms of service providers will play a more definite role in the future. This could in turn boost new business models and new forms of revenue sharing through the advent of new services.

The content services that will be provided by UMTS will evolve from the services provided by GSM. These services offered by UMTS can be classified as "mobile Internet", where display text or other contents are optimized for the limited graphical capabilities of the mobile devices, along with the location awareness functionality. A range of services have been suggested, for example: *Images and video:* streaming pictures and video, entertainment, games, lottery, gambling, and video-conferencing, *Short range and Location based services:* Information of nearby restaurants, shops and special offers, *Context based services, or push:* Notification of interesting sports event if you are in front of the TV, advertising if you are close to the advertiser, Simultaneous transfer of speech, data, text, pictures, audio and video, High-speed, mobile access to the Internet, Customized infotainment, *Travel information:* congested roads, flight departures, and if you get lost, find your current location etc.

When mobile users are roaming (moving from one area or cell to another) the handset must attach itself to the closest base-station. When one base-station hands over the user to another base-station, multiple transactions must be made. If the user is roaming among base-stations belonging to the subscribed operator, this is only a technical issue. But when roaming implies other operators, national or international, the picture becomes more complicated. The vision of UMTS presupposes that it will be possible for users to roam freely with their handsets and have affordable access to equivalent products and services globally by end-users utilizing the most proper network despite which operator it belongs to. Today, technically and efficient roaming exists for the GSM standard throughout Europe, parts of Africa, Asia and South America. However, it is far from global: for instance, roaming to and from North America and Japan requires special measures and users may even incur custom duties and taxes relating to the use of their handsets abroad. Moreover, the complexity and non-transparency of pricing for roaming services have been problematic for 2G or 2.5G users [2]-[8].

Some related works on UMTS can be seen in [2]-[7]. The paper is organized as follows. Section 2 discusses the features of UMTS, section 3 is the architecture details, section 4 is the

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_98,

simulation of two UMTS scenarios where we use different network topologies and section 5 is the conclusion.

II. UMTS Features

A. Fast access

UMTS is better than 2G or 2.5 G cellular technology for its potential to support 2 Mbit/s data rates. This capability, together with the built-in IP support, merges powerfully to allow interactive multimedia services and new wideband multimedia-rich applications, such as video telephony, video conferencing, network video gaming etc.

B. Packet transmission and data rate on demand

Current cellular systems use circuit switched technology for data. UMTS combines packet and circuit data transmission with the advantage of virtual connectivity to the network at all times and alternative ways of billing. For example, pay-per-bit option, to pay per session, to pay flat rate etc as demanded by many upcoming data services. The operation of the system can be made much cheaper as UMTS is also being designed to offer data rate on demand which can work in combination with packet data transmission.

C. Friendly and consistent service environment

UMTS services are based on having the user face common capabilities throughout all UMTS user and radio environments. When roaming from the user's network to other UMTS operators, a personal user will experience a consistent set of services and this would give a feeling or sense of his or her home network (termed as "Virtual Home Environment" or VHE). VHE will ensure the delivery of the service provider's full environment. VHE will enable terminals to handshake and agree on some functionalities with the visited network, and "home network like" services will be presented with full security and transparency across a mix of access and core networks.

D. Mobility and Coverage

UMTS is implemented as a global system, combining both terrestrial and satellite communication network components. Multi-mode terminals operating also via 2.5 or 3G systems will further extend the reach of many UMTS services. With these terminals a user will be able to move around and roam from a private network into a micro-cellular public network, then into a wide area macro-cellular network (like a 2.5 or 3G network), and then to a satellite mobile network, with minimal break in communication and packet loss.

E. Radio Technology for all environments

The UMTS radio interface UTRA (UMTS Terrestrial Radio Air interface) can help operation with high spectral efficiency and service quality. In practical implementations, UMTS terminals might be unable to function at the peak data rates at all times, and in remote or heavily congested areas system services might only support lower rates due to radio propagation constraints or other economic reasons. In order to enable subscribers to always use their terminal, services will be adaptive to different data rate availability and other Quality of Service (QoS) parameters. In the early stages of deployment, UMTS coverage can be limited. UMTS can enable roaming with other networks like GSM system operated by the same operator or with other GSM or 3G systems of other operators, including UMTS compatible satellites.

F. UMTS Services global availability through Satellite

Satellite technology can give enhanced global coverage and services and it is expected to play an important role for UMTS world-wide coverage. UMTS is being standardized to ensure an efficient and effective roaming and handover between satellite and terrestrial networks.

G. Spectrum for UMTS

WRC-92 identified the frequency bands 1885-2025 MHz and 2110-2200 MHz for future IMT-2000 systems, with the bands 1980-2010 MHz and 2170-2200 MHz intended for the satellite part of these systems. These bands can be made available for UMTS licenses to enable operators for network deployments. Wideband Code-Division Multiple-Access (W-CDMA) is one of the main technologies for the implementation of third-generation (3G) cellular systems. In W-CDMA interface different users can simultaneously transmit at different data rates and data rates can even vary in time. UMTS networks need to support all current 2G/3G services [5]-[7].

III. UMTS Architecture

A UMTS network consists of three domains that interact with each other, namely – Core Network (CN), UMTS Terrestrial Radio Access Network (UTRAN) and User Equipment (UE) [8]. The main function of the core network is to provide switching, routing and transit for user traffic. Core network also contains the databases and network management functions. The basic Core Network architecture for UMTS is based on GSM network with GPRS. All equipment has to be modified for UMTS operation and services. The UTRAN provides the air interface access method for User Equipment. Base Station is referred as Node-B and control equipment for Node-B's is called Radio Network Controller (RNC).

It is important for a network to know the approximate location in order to be able to contact the user equipment. Some of the list of system areas from largest to smallest – UMTS systems (including satellite), Public Land Mobile Network (PLMN), MSC/VLR (Mobile Switching Center)/(Visitor Location Register) or SGSN (Serving GPRS Support Node), Location Area, Routing Area (PS domain), UTRAN Registration Area (PS domain), Cell and Sub cell etc.

A. Core Network

The Core Network consists of circuit switched and packet switched domains. Some of the circuit switched elements are

Mobile services Switching Centre (MSC), Visitor location register (VLR) and Gateway MSC. Packet switched elements are Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). Network elements like EIR (Equipment Identity Register), HLR (Home Location Register), VLR (Visitor Location Register) and AUC (Authentication Center) are shared by both domains. The Asynchronous Transfer Mode (ATM) is defined for UMTS core transmission. ATM Adaptation Layer type 2 (AAL2) handles circuit switched connection and packet connection protocol AAL5 is designed for data delivery. The architecture of the Core Network may change when new services and features are introduced. Number Portability Database (NPDB) will be used to enable user to change the network while keeping their old phone number. Gateway Location Register (GLR) may be used to optimize the subscriber handling between network boundaries.

B. Radio Access

Wide band CDMA technology was opted for UTRAN air interface. UMTS WCDMA uses Direct Sequence CDMA where user data is multiplied with quasi-random bits derived from WCDMA spreading codes. In UMTS, in addition to channelization, codes are used for synchronization and scrambling. WCDMA has two basic modes of operation: Frequency Division Duplex (FDD) and Time Division Duplex (TDD). The functions of Node-B are stated as follows: Air interface Transmission/Reception, Modulation/Demodulation, CDMA Physical Channel coding, Micro Diversity, Error Handing and Closed loop power control. The functions of RNC are – Radio Resource Control, Admission Control, Channel Allocation, Power Control Settings, Handover Control, Macro Diversity, Ciphering, Segmentation/Reassembly, Broadcast Signaling and Open Loop Power Control.

C. User Equipment

The UMTS standard does not hinder or cripple the functionality of the User Equipment. Terminals work as an air interface counter part for Node-B and have many different types of identities which are taken directly from GSM specifications. They are International Mobile Subscriber Identity (IMSI), Temporary Mobile Subscriber Identity (TMSI), Packet Temporary Mobile Subscriber Identity (P-TMSI), Temporary Logical Link Identity (TLLI), Mobile station ISDN (MSISDN), International Mobile Station Equipment Identity (IMEI), International Mobile Station Equipment Identity and Software Number (IMEISV).

UMTS mobile station (MS) can operate in one of three modes: *PS/CS (Packet Switched/Circuit Switched) mode of operation:* The MS is attached to both the PS domain and CS domain, and the MS is capable of simultaneously operating PS services and CS services. *PS mode of operation:* The MS is attached to the PS domain only and may only operate services of the PS domain. However, this does not prevent CS-like services to be offered over the PS domain (like VoIP). *CS mode of operation:* The MS is attached to the CS domain only and may only operate services of the CS domain.

UMTS IC card has similar physical characteristics as GSM SIM card. It has various functions as follows – support of one User Service Identity Module (USIM) application (optionally more that one), support of one or more user profile on the USIM, Update USIM specific information over the air, security functions, user authentication, and optional inclusion of payment methods and optional secure downloading of new applications [8]-[9].

IV. UMTS HANDOVER

There are following categories of UMTS handover (also referred to as handoff) that can happen:

A. Hard Handover

Hard handover means that all the old radio links in the UE are removed before the new radio links are established. Hard handover can be seamless or non-seamless. Seamless hard handover means that the handover is not perceptible to the user. In practice a handover that requires a change of the carrier frequency, known as inter-frequency handover is always performed as hard handover.

B. Soft Handover

Soft handover means that the radio links are added and removed in a way that the UE always keeps at least one radio link to the UTRAN. Soft handover is performed by means of macro diversity, which refers to the condition that several radio links are active at the same time. Normally soft handover can be used when cells operated on the same frequency are changed.

C. Softer handover

Softer handover is a special case of soft handover where the radio links that are added and removed belong to the same Node B (i.e. the site of co-located base stations from which several sector/cells are served). In softer handover, macro diversity with maximum ratio combining can be performed in the Node B, whereas generally in soft handover on the downlink, macro diversity with selection combining is applied. Active Set is defined as the set of Node-Bs the UE is simultaneously connected to (i.e., the UMTS Terrestrial Radio Access cells currently assigning a downlink DPCH (Downlink Dedicated Physical Channel) to the UE).

V. SIMULATION DETAILS

A. Scenario 1 (Soft and hard handovers between UMTS networks)

We used existing simulation models and only verified the existing results and no modifications were done to the model. Our aim is to conduct a performance comparison between soft and hard handovers between UMTS networks (with different towers). Consider fig. 1. The configurations done are given below. In these scenarios, the two UEs (User Equipment) move around and make zigzag (back and forth) transition between two Node-Bs (UMTS towers) creating multiple handovers from one cell to the other. Continuous upload of large FTP files to the server was done by the two UEs during the simulation, which makes the traffic move in the uplink direction in the UMTS part of the network. UMTS QoS 3 is used for the traffic associated with FTP.

The configurations done and the network topology of both scenarios are same except that in hard handover case, the soft handovers are disabled at the RNC (Radio Network Controller), so that all the handovers done are hard handovers. To study the soft handover case, they are all configured as soft handovers.

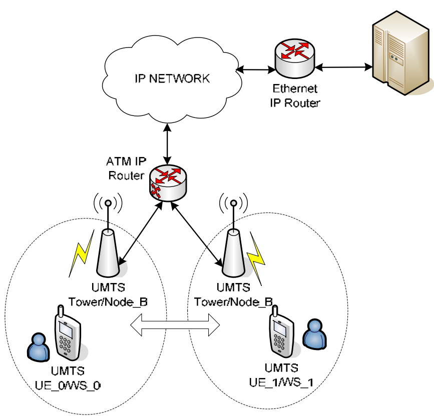


Fig. 1. Network Scenario 1

In this study, we observed the results for two statistics that are used for QoS 3 traffic: The FTP Upload response time and the Uplink Transmission Power of the physical channels in the UE (User Equipment). The ASC (Access Service Class) 6 is selected here for UMTS.

ASC classes are defined at the RNC node. There are a maximum of 8 access service classes available. As we investigated and compared application response times from the two scenarios, it showed that there is not much difference between hard and soft handovers in terms of their effect and impacts on application response times.

On the other hand, for both UMTS user equipments, soft handover produced better results in terms of uplink transmission power. While observing the uplink transmission graphs, the peaks or highs indicate handovers. The peaks are smoother during soft handovers. The UEs go away from their Node-Bs, which increases the transmission power and after handovers they are connected to the nearest Node-B resulting in lower transmission power. The advantage of soft handovers over hard handovers can be seen as follows. When the line graph of these two handovers is compared, the UE transmission power reaches to higher values for the hard handovers compared to the soft handovers. The difference of transmission power is around 3 dBm roughly. Thus soft handovers enable the UEs to perform the handovers at power levels that are low, which is an added benefit for other UEs around when we consider the issue of interference.

To achieve equal performance benefits with lower power, soft handovers benefit from an idea that can be termed as "soft handover gain". As a radio bearer is aided by more than one radio links during the soft handovers, the network can lower the quality (BLER, Block Error Rate) of these links below the requested level as long as their combined or integrated quality is enough to provide the needed or requested quality for the radio bearer. Thus the transmission power can be maintained at lower levels during soft handovers, as a result of lowering the radio link quality, while still providing the same quality to higher layer as the results show us. Have a look at fig. 2 to 5.

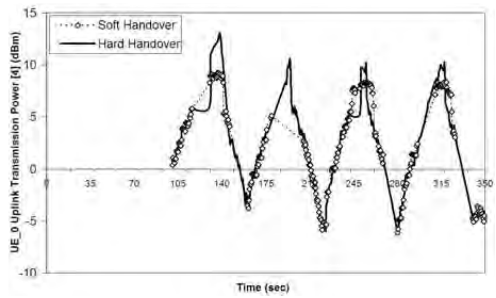


Fig. 2. Uplink transmission power (UE 0)

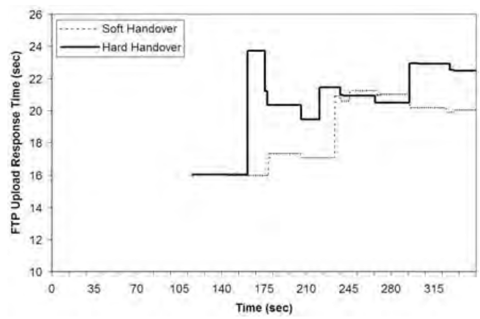


Fig. 3. FTP Upload response time

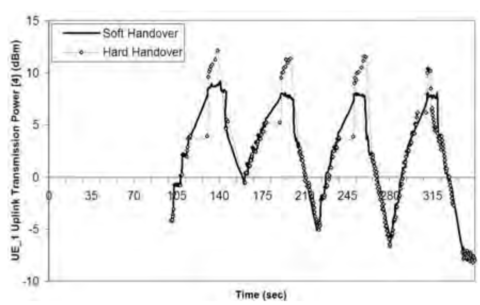


Fig. 4. Uplink transmission power (UE 1)

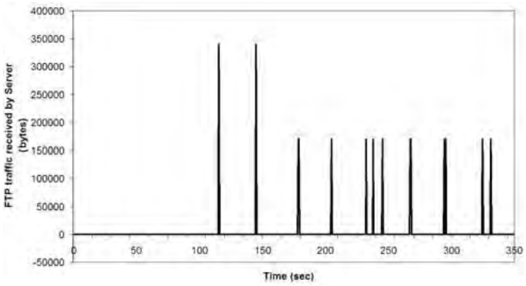


Fig. 5. FTP traffic received by Server (bytes/sec)

B. Scenario 2 (Softer and hard handover between three sectors)

Our aim is to demonstrate the softer and hard handover process that happens to mobile user equipment while it is around a UMTS Node B with three sectors or cells. We also want to study the user equipment's (UE) transmission power between the softer and hard handover scenarios. Softer Handover takes place when the UE is positioned in an area where sectors are overlapped. Hard handovers take place when the UE is moving from one sector to another.

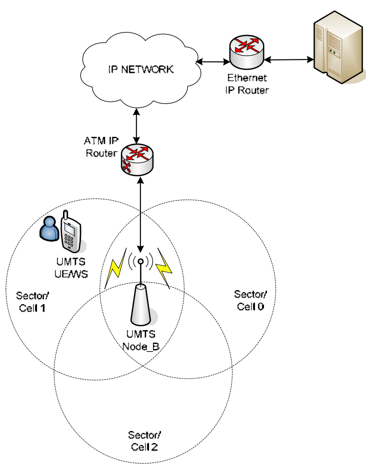


Fig. 6. Network Scenario 2

Consider fig.6. The configuration used is as follows. The Node B has three directional antennas that cover an angle of 160 degrees. The sectors/cells are positioned every 120 degrees starting from 0 degrees and so it would be 0, 120, and 240 degrees. Thus an inter sector/cell overlap of 40 degrees is achieved between neighboring cells.

In these simulation scenarios, the UEs follow its pre-defined path around the Node B causing multiple handovers as it moves from one sector to another sector. Using UMTS QoS 3 service level, the UE also uploads files through FTP to a remote server at a distant network.

Some simulation parameters or settings are as follows: The RNC (Radio Network Controller) UMTS handover parameter – *Soft Handover* is set to "Supported" in the softer handover scenario and to "Not Supported" in the hard handover scenario. The configurations done and the network topology of both scenarios are same except that in hard handover case, the soft handovers are disabled at the RNC (Radio Network Controller), so that all the handovers done are hard handovers. To study the soft handover case, they are all configured as soft handovers as explained above. The *Active Set Size* parameter is set to 2.

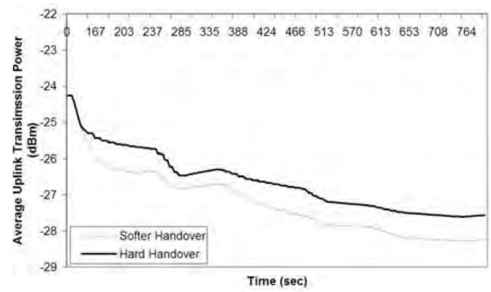


Fig. 7. Uplink transmission power

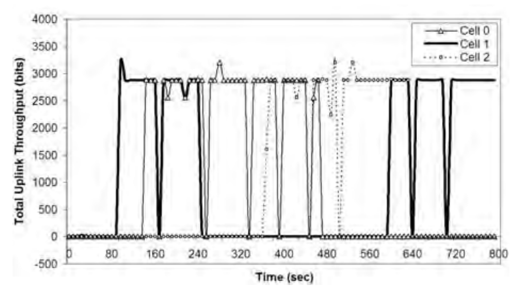


Fig. 8. UMTS Cell/Sector Uplink throughput

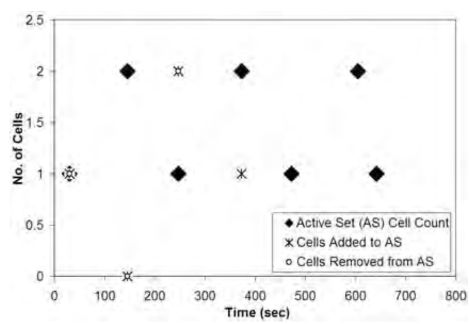


Fig. 9. UE active set details

Here we discuss some statistics and to illustrate the softer handover process. These statistics show the number of cells in user equipment's Active Set, showing which cells are included and discarded from that set during the simulation performed.

As the UE travels around the Node B, it notes the pilot channel signal strength that emanates from all the three sectors. These signal strength measurements allows it to change its active set.

Let's have a look at the softer handover events that occurred during simulation. The UE path is from sector 1 to sector 0, followed by sector 0 to sector 2 and finally from sector 2 to sector 1. At the start of simulation, based on the strongest signal from sector 1, this sector is included in the initial active set. The UE starts moving at 115 sec. After some time when it reaches the edge of sector 0 at 175 sec, adding sector 0 into the active set. The UE is now in softer-handover between sectors 0 and 1 while it remains in this position for around 60 seconds. Now the UE traverses toward the middle of sector 0 at 235 sec (remaining in sector 0 for 60 seconds) and thus UE leaves sector 1 deleting it from its active set at 255 sec. At 375 sec the UE moves to sector 2. While UE is in the overlapped area between sectors 1 and 2, it gets into softer-handover state once again at 369 sec. The UE continue moving along its path to the middle portion of sector 2, deleting or removing sector 0 and ending its softer-handover state at 412 sec. In the last traversal, the UE comes back to sector 1 and enters into a softer handover state when it lies on the area shared by both sectors (1 and 2) at 605 sec. Then the UE goes to the center of sector 1, dropping sector 2 from its active set at 650 sec.

Thus Softer Handover takes place when the UE is positioned in an area where sectors are overlapped. Hard handovers take place when the UE is moving from one sector to another. In both scenarios, the results for several statistic groups are collected like: UE active set details, Node Bs throughput, UE RLC/MAC uplink transmission power etc.

Let's have a look at the hard handover events that occurred during simulation. At the start of simulation, based on the strongest signal from sector 1, this sector is included in the initial active set. UE starts moving at 115 sec. After some time when it reaches the edge of sector 0 at 175 sec, a hard handover happens. Sector 0 gets added into the active set while sector 1 is taken off. At 355 sec, the UE traverses to sector 2 and a hard handover happens from sector 0 to sector 2 at 369 sec. Finally the UE returns to sector 1 and performing another hard handover state, but now from sector 2 to 1 at 605 sec. Then the UE goes to the middle of sector 1 and remains there at 650 sec.

As we compared the UE transmission power between both softer and hard handover scenarios, we found that the UE in the softer handover scenario uses less transmission power, thus reducing the cell interference. But on the other hand during softer-handovers the UE is taking more network resources.

Note that during the softer handover scenario, the UMTS Node B is getting traffic through two sectors at the same time. But during the hard handover case, only one sector receives traffic. See fig.7, 8 and 9 to help understand the scenarios fully.

VI. Conclusion

The Universal Mobile Telecommunications System (UMTS) is one of the third-generation mobile telecommunications technologies, which is currently worked upon to be developed into a 4G technology. The technology was proposed by 3GPP (3rd Generation Partnership Project) and is part of the global International Telecommunication Union IMT-2000 (International Mobile Telecommunications-2000) standard. 3GPP specifications are based on evolved Global System for Mobile Communications (GSM) specifications. A study is done on the different scenarios of UMTS network and its performance evaluation is done when the mobile node is on the move. We were able to look into two network scenarios where different types of handovers were investigated and the performance graphs studied. This would help to understand the issues in soft and hard handover scenarios if one were to look into the details of UMTS technology.

References

[1] EU, Decision No 128/1999/EC of the European parliament and of the council, of the 14 December 1998, on the coordinated introduction of a third-generation mobile and wireless communications system (UMTS) in the Community, 1998, available [online]: http://europa.eu.int/eur-lex/pri/en/oj/dat/1999/l_017/ l_01719990122en00010007.pdf

[2] Saravanan, N.; Sreenivasulu, N.; Jayaram, D.; Chockalingam, A., "Design and Performance Evaluation of an Inter-System Handover Algorithm in UMTS/GSM Networks", Proceedings of TENCON 2005, 2005, pp.1-6.

[3] Oudelaar, J., "Evolution towards UMTS", Proceedings of 5th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 1994. Wireless Networks - Catching the Mobile Future, 1994, pp.852-856.

[4] Bosch, P.; Mullender, S.; Samuel, S., "Mobility in UMTS packet IP networks", Proceedings of Wireless Communications and Networking Conference, 2006, pp. 545 - 550.

[5] Pereira, A.; Gameiro, A., "Simulation-based Evaluation of Spectrum Opportunities in UMTS Cellular Networks", Proceedings of Mobile and Wireless Communications Summit, 2007, pp.1-5.

[6] Website: http://www.cellular.co.za/umts_technical.htm, Retrieved March 15, 2009.

[7] Website: http://www.umtsworld.com/technology/wcdma.htm, Retrieved March 20, 2009.

[8] Petter Nielsen, Introduction of the third generation of mobile communication system in Norway, Department of informatics, University of Oslo, Norway. Website: http://heim.ifi.uio.no/~pnielsen/PHD/Ni.doc, Retrieved March 25, 2009.

[9] UMTS World, Overview of The Universal Mobile Telecommunication System, Online: http://www.umtsworld.com/ technology/overview.htm, Retrieved February 2009.

Reduction of Energy Consumption in WSN using an Approach Geocast

Harilton da S. Araújo
Technology Sciences Center
Master in Applied Computer
University of Fortaleza
Fortaleza – Ceará – Brasil
harilton@edu.unifor.br

Wagner L. T. de Castro
Technology Sciences Center
Computer Science Dept.
University of Fortaleza
Fortaleza – Ceará – Brasil
wagner.ltc@edu.unifor.br

Raimir Holanda Filho
Technology Sciences Center
Master in Applied Computer
University of Fortaleza
Fortaleza – Ceará – Brasil
raimir@unifor.br

***Abstract* – Wireless sensor networks have deserved special attention in the last years due to their intrinsic characteristics that become them attractive in various applications, such as in the industrial activities, in the environment monitoring and security, among others. In general, this kind of network can be utilized as powerful tools in order to monitor and, eventually, control an environment. The routing protocols, for the WSN, must have self-configuration characteristics aimed at discovering which is the best way for transferring the information, with delivery assurance and with minimum energy consumption, among the nodes that compose the network. This article proposes a modification of the Directed Diffusion routing protocol so as to reduce the energy consumption in the network when there is an occurrence of failures. The proposal utilizes a Geocast approach to repair broken paths by constructing a new routing tree. The performed simulations demonstrated that our proposal is more efficient from the energy viewpoint.**

Keywords - wireless sensor; networks; routing; geocast

I. INTRODUCTION

A network of sensors is composed of a large number of nodes that are utilized as per request at an area of interest. Each node has one or more sensors, in addition to the capacity for processing, storage and communication.

The Wireless Sensor Networks have aroused the interest of the scientific community due to their applicability that reaches various areas such as the military, environmental, medical and industrial ones. Those networks differ in several aspects from the traditional networks. Some characteristics are: they are composed of a large number of sensor nodes; the sensors have high energy limitations, and low processing and memory capacity. In addition to that, some WSN applications require characteristics of self-organization, being able to adjust themselves, in an autonomous way, to the possible structural changes due to external interventions such as modifications of the topology caused by failures, inclusion of nodes or as per a request submitted by an external entity (user).

The purpose of the WSN is to collect data of a region being covered by sensors and to allow the extraction of that data, by an external entity, through the sink node.

The energy consumption is the main requisite of the network and it affects all of the phases of its life cycle, not being dependant upon the application. The communications in WSN consume more energy than the processing and the sensor coverage performed by the network nodes. This characteristic requires routing protocols that make possible that the sensor nodes communicate themselves in an efficient and effective way with minimum energy consumption.

The routing protocols for sensor networks must have self-configuration characteristics in order to discover which the best way is for transferring the information from a node to the other one, with delivery assurance and minimum energy consumption, having the purpose of extend the lifetime of the network. If a node fails, through which a given information flows, a new routing will have to be made in order that the information to be collected is able to reach the destination node. The communications, among the network nodes, must be made so as the energy consumption is optimized. In this sense, some protocols have the essential role of increasing the time of the network useful life.

This article proposes a modification of the Directed Diffusion routing protocol, having the purpose of reducing the energy consumption in the network by means of focusing on the presence of failures. A Geocast approach was used in order to reduce the number of sent and received control messages. That reduction was reached through the construction of a new routing tree being limited and directed towards a specific region of the network, focusing the occurrence of failures.

The article is organized in the following way: the Section II presents the linked works. The Section III describes the modifications that we made in the Directed Diffusion protocol. The evaluation of the algorithm is discussed on the Section IV, followed by conclusions on the Section V.

II. RELATED WORKS

A proposal for optimization of the energy consumption is analyzed on [1]. The conception of this proposal is to put some specific nodes in the sleep (disconnected) mode in order

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_99,

to save energy, while keeping the connectivity of the network thus assuring the communications.

The EAR (Energy Aware Routing) protocol is described on [2]. The EAR protocol is a reactive one that looks for paths of minimum energy consumption through the network. The basic operation consists of the occasional use of a set of paths chosen through a probability function (which is dependant upon the energy consumption of each path). In this way, it avoids that the best path has its energy depleted, thus increasing the network utilization time. It is assumed that each node has an address and information about its localization.

An algorithm is proposed on [3], which utilizes GPS in order to restrict the flooding area of the packages. This algorithm uses two approaches in order to determine which nodes must propagate the control messages. The first one is determined based on the position and the size of the area. The second one is established according to the distance and the localization of the destination.

A routing protocol, that limits the research aimed at looking for new routes in ad hoc networks, is described on [4]. It is utilized, in that proposal, an approach based on flooding and on the localization of the nodes in order to limit the flooding area. When sending a message, the emitter node defines an "expected zone" of the destination node. If any node, that receives the message, is outside that zone, it is discarded. The purpose is to reduce the cost of the retransmissions of each message through the whole network.

Alternative definitions of requisition zones, being aware of the localization, are analyzed on [5]. Those definitions have the purpose of optimizing the performance by delimiting the areas of interest. In order to do that, the areas are approached based on the rectangular, circular or conic format.

An algorithm for data dissemination (Directed Diffusion) is described on [6], where it is proposed a routing scheme being centered on data, there not being semantics of directing. It also attempts to take care of dynamic networks through a negotiation scheme with dissemination of interests and reinforcements of paths, allowing the network to converge in the presence of any topologic modification. In addition to that, the Directed Diffusion utilizes, as metrics for the selection of the route, that one that has the lower delay, which generally leads to the route with a lower quantity of hops. A disadvantage of this approach is the high cost of communications for repairs of paths when there is occurrence of failures, due to the need of periodically performing a flooding in the network in order to reinforce other paths.

We present, on this article, a proposal of routing protocol for wireless sensor networks. The proposal is implemented as a modification of the Directed Diffusion protocol, and it differs from the solutions presented in the literature for the fact that it uses an approach to minimize power consumption during recovery from failures. Localization information is utilized in order to reduce the energy consumption. The algorithm is described in detail on the following Section.

III. Proposal for an Energy Aware Protocol

We propose, on this article, an extension towards an existing data centric routing protocol, Directed Diffusion [6].

The main purpose of our proposal is to develop a protocol being tolerant to failures and having a minimum of energy consumption. In order to simultaneously reach those requisites, an approach, based on the Geocast routing method, was utilized in our algorithm. The approach consists in repairing broken paths by constructing a new, limited and directed routing tree. The approach used in the proposed routing algorithm is described as follows:

A. Limited and Directed Geocast Approach

Based on [5], [7] and [8], the limited and directed geocast approach was created in order to perform the construction of a new routing tree being limited and directed towards a specific region of the network denominated the geocast region. In order to do that, it becomes necessary the use of geographical localization information regarding the nodes, which allows to direct the routing of the messages. The approach, as adopted on this article, restricts, in the rectangular format, the region that contains the incidence area of events. At the time of receiving a message, each node verifies whether it belongs to the geocast region. If it belongs, the node sends the message again to its neighbors; in the contrary case, the node disregards the control message. It is assumed that all of the nodes know their geographical positions.

The scheme that restricts the geocast region is similar to the optimization scheme as quoted on [7]. This approach restricts the flooding area by the straight lines r_2,r_3,r_4,*e* r_5. It assumes that the node A_1 (Sink) knows the localization (X_{A2},Y_{A2}) of the node A_2 (Source). Being r_1 the straight line that passes through the points A_1 and A_2. The distance, between the points A_1 and is represented by $D_{A1,A2}$.*P* is the mid-point between A_1 and A_2. The radios *R* is the distance utilized to draw the parallel straight lines(r_2*e*r_3)and the perpendicular straight lines (r_4*e*r_5) towards the straight line. The distance between and is represented by:

$$D_{B1,B2} = 2R + D_{A1,A2} \qquad (1)$$

If the distance between and is represented by , then the distance between and P is indicated by:

$$D_{B1,P} = \frac{(2R + D_{A1,A2})}{2} \qquad (2)$$

The figure 1 shows an example of the restricted geocast area.

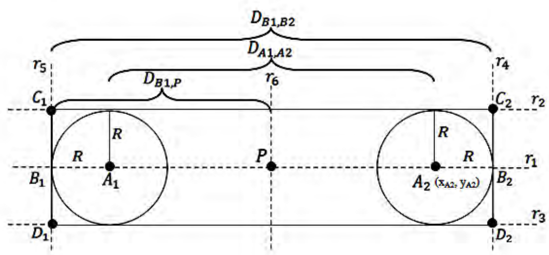


Figure 1. Restriction scheme in the geocast region

A field, denominated (number of hops), was created in order to calculate the number of hops as from the "sink" node. The value of the radios is calculated as being the distance between the sink node and the most distant node that could be reached with "X" hops. Where "X" is defined by the user. The radios construction message is sent by utilizing a determined integer number that will be stored in the field, which will be decreased one unit for each retransmission and when reaching zero, the radios construction message will be discarded and a message, informing the geographical position of those nodes, is sent to A_1 (Sink) in order to determine which the most distant node is. In that way, the radios construction message will have met its target, that is to say, reaching the nodes by meeting the number of hops as predetermined in the configuration phase. The size of the geocast region is calculated on the basis of and $D_{A1,A2}$. The reduction of the area, for the construction of a new routing tree, gets advantages as the number of control messages to be interchanged will be reduced. The lower the is, the lower the size of the broadcast zone will be.

Following that restriction scheme of the area, the nodes must propagate the control messages only for the nodes that belong to the geocast region as it is shown on the figure 2.

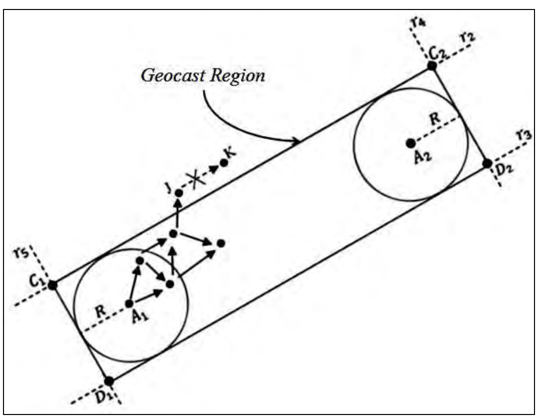


Figure 2. Propagation of messages

The propagation of messages performs the following steps:

- The information, needed to create the geocast region is sent on the control message;
- At the time of receiving a control message, the node verifies whether it belongs to the restricted area;
- If it belongs, it propagates the message to the neighboring nodes, in the contrary case, it discards the message as it is the case of the J and K nodes.

B. *Proposed Routing Algorithm*

Our routing algorithm is implemented in four phases:

1) *Configuration Phase*

In each run of the protocol, the sink node performs floodings with messages of interest for its neighbors. The messages of interest describe the task by means of an attribute-value pair [6]. When a node receives an interest, it verifies whether the interest exists in the cache. If there is no linkage with any of the various stored interests, the node creates a new entry of interest. Each entry in the cache of the node has a gradient that points the interest for the neighbor that sent it. By utilizing interests and gradients, paths are established among the sink node and the source nodes.

Upon receiving an interest, a node resends this interest by utilizing a gradient. For their neighbors, the interest appears as having been originated by the node that is the emitter of the message, although it may come from another distant node. In this way, this protocol performs just local interactions for the dissemination of the interests.

2) *Phase of Route Establishment*

When the interest arrives at an appropriate region, through the exploratory messages, one or more sensors are enabled, thus becoming sources. The source nodes send data to the neighbors that have a gradient. The first data are called exploratory data. Within the directed diffusion, the nodes resend the first exploratory data, that were received, up to arriving at the sink node, favoring the paths of low latency. According to what was stated on [9], the route, that has low latency, is not an effective metrics in WSN. In our protocol, each node inserts, in the exploratory message, the number of hops and the sum of the residual energies of each node, which composes the path between the source and the destination. The data message is immediately sent to its neighbors. The exploratory messages are stored at a local cache in order that the selection of routes is performed. A timer is created in each node of the network and it is utilized after the reception of the exploratory messages. The use of the timer is appropriate in order to allow the exploratory messages, from the most distant nodes, to arrive and that they are considered for the selection of the path.

The selection of the routes is performed by utilizing the same mechanism as adopted on [10], which allows the nodes to locally determine the route along which the data will be conveyed without incurring in high cost related to the knowledge of the whole network topology. The selection process of routes can be represented as follows. Being a directed graph, where is a set of vertexes (sensor nodes) and is the set of pairs , not ordered of elements in (connections). In case that a message is conveyed through a path $C = (\ldots)$ in a graph , where , ... , are vertexes and , ... ,) are edges, then each node in loses some of its energy linked to the cost for sending messages. Being the cost for sending the message, , the initial energy, and the residual energy of after sending a message. Hence, for . Consequently, the energy of the path is indicated by , where:

$$E_C = \sum_{i=1}^{k-1} R_{v_i} \qquad (3)$$

The rule, as implemented in the protocol, does not choose just the path that has the highest value of available residual energy for all of the nodes, but also the number of nodes that compose the path. As the solution utilizes just localized interactions, the reason is used for getting the best path. The reason is got by means of the division between the sum of the residual energy ( ) and the number of hops that compose the path ( ), according to what the equation shows .

$$\lambda = \frac{E_C}{k-1} \qquad (4)$$

Consequently, it is considered that the best path to be chosen is that one which, amongst all of the available paths of a given pair of vertexes of origin and destination, gets the highest value corresponding to the reason .

On the basis of the exploratory messages and the mechanism for selection of routes, the best path is selected, which means that it is the path that has the highest residual energy with the lowest number of hops.

3) Phase of Data Communication

The phase of data communication occurs by utilizing a gradient between the receiver node and the node that sent the message. The gradient is established by the exploratory message. As soon as the source node starts sending its data, the messages are conveyed by the intermediary nodes following the direction as pointed out by the gradient as predetermined in the prior phase. Each one of the intermediary nodes conveys the package of data up to arriving at the sink node.

4) Reconstruction Phase of Paths

In case of failure occurrence regarding a path between the source node and the sink node, an alternative route must be established. In order to do it, basically, the Directed Diffusion protocol restarts the reinforcement mechanism to search for other paths. However, the repair scheme of paths in the Directed Diffusion has a high cost as regards to the energy consumption as it requires a flooding of the network to reinforce other paths.

In the case of occurrence of failures in the nodes that complete the stretches, our algorithm utilizes a Geocast approach instead of performing a flooding of the network for reinforcement of other paths. That approach has the purpose of repairing broken paths by constructing a routing tree being limited and directed towards a specific region of the network in which the set of source nodes is located, thus minimizing the number of control messages being sent and received. The approach restricts, in the rectangular format, the area where the new routing tree will be constructed. At the time of receiving the package, each node verifies whether it belongs to the geocast region. If the package belongs, the node sends the message again to its neighbors; in the contrary case, the node disregards the control message.

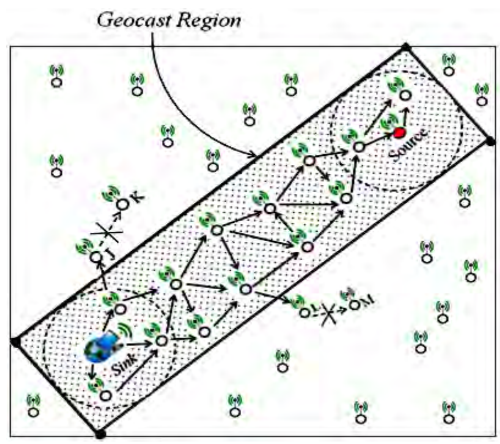


Figure 3. Reconstruction of paths

At the time of the definition regarding the geocast region, the protocol disregards the node that failed, in order that this node does not compose the new routing tree that will be created. At the end of the transmission, referred to the event that was interrupted due to failure of the route, the protocol restarts the configuration phase. The failures in the routes are identified when the sink node stop receiving the events associated to the interest served.

The approach, as utilized in that algorithm, assures the quick reconstruction of paths in the presence of failures and the reduction of the energy consumption as it avoids the flooding in the whole network for the reconstruction of new routes. The main difference, between the approach as utilized in the Directed Diffusion and what is being proposed on this article, is the construction of the routing tree by means of the occurrence of failures. In the first approach, the broadcast for the construction of the tree has maximum reaching, that is to say, it reaches all of the network nodes. According to the second approach, the limited and directed construction of the routing tree is performed, having the purpose of reaching just the nodes that are close to the occurrence of events.

IV. Evaluation of Performance

In order to evaluate the performance of our proposal, we utilized the Sinalgo simulator [11]. Sinalgo is a framework, written in Java, that allows the simulation of wireless networks and abstracting itself from the lower layers of the network stack.

A. *Simulation Scenario and Metrics Utilized*

The simulation scenario was constructed in order to allow a didactic understanding of the proposed algorithm. The topology of our scenario allows that each node communicates with up to 8 neighboring nodes. There are 121 nodes, in an 11 x 11 network topology.

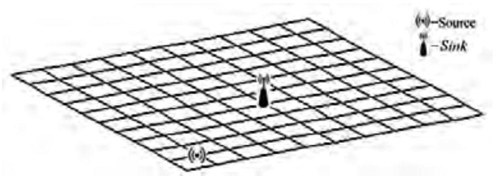


Figure 4. Topology of the Scenario

There is a data flow that consists of a source and a destination (Sink). The source node generates an event and it is located as it is shown on the figure 4. The total simulation time was 500 seconds and we repeated the simulation three times. The messages were simulated with packages of 500 bytes. The duration of the interest was 150 seconds. The energy dissipation pattern, that was adopted, was the same as the one utilized in [12]. The initial energy of the nodes was adjusted to 1,5J. Three metrics were chosen in order to evaluate the performance of our proposal as compared with the Directed Diffusion protocol: network energy map, energy consumption and cost of receiving messages. The network energy map shows the residual energy of each sensor node. The consumption of energy measures the total energy rate that is available at all of the network nodes. The cost of receiving messages measures how many messages are delivered for each unit of energy (J) used by all network nodes.

B. *Results Got*

The energetic distributions of the network along the time *t*=450s are shown on the figures 5 and 6. Analyzing the surface, we observe the following results: on the Directed Diffusion (Figure 5), the energy is consumed at all of the network nodes and because of that, it can be observed that the mesh is positioned at a lower level than the one of our proposal (Figure 6), that occurs due to the periodical broadcasts for the reconstruction of routes at the Directed Diffusion, while in our proposal the broadcasts are limited to the geocast region and they are directed to the region of interest.

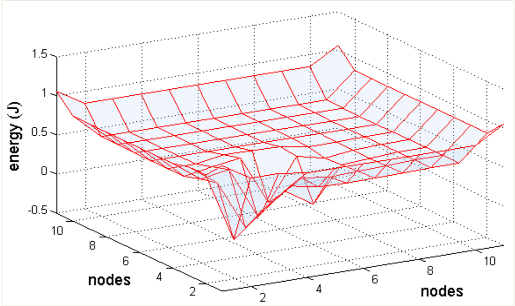


Figure 5. Energy map (Directed Diffusion)

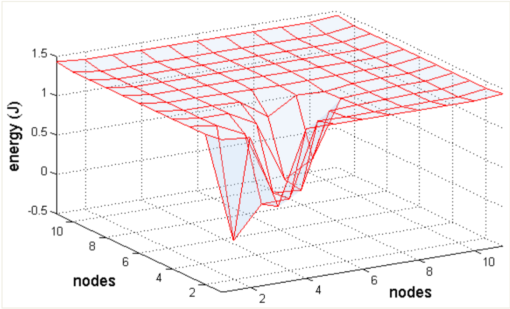


Figure 6. Energy map (Directed Diffusion + Geocast Approach)

The depressions, as shown on the figures 5 and 6, refer to the region between the sink node and the source node. It is perceived that as regards to the initial energy status of the nodes (1,5J) and not regarding the positioning of the mesh, our proposal shows a less accentuated depression as a result from the mechanism for selection of routes that is utilized, which propitiates a balancing of the energy consumption.

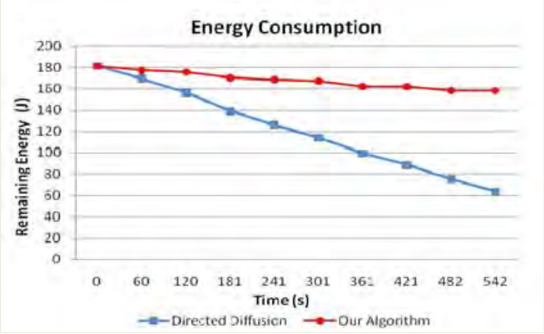


Figure 7. Energy consumption

The Figure 7 shows that the energy consumption, in the whole network, is reduced by utilizing our proposal. This is got due to the use of the geocast approach regarding the mechanism for selection of routes, which chooses the path that has the highest residual energy with the lowest number of hops.

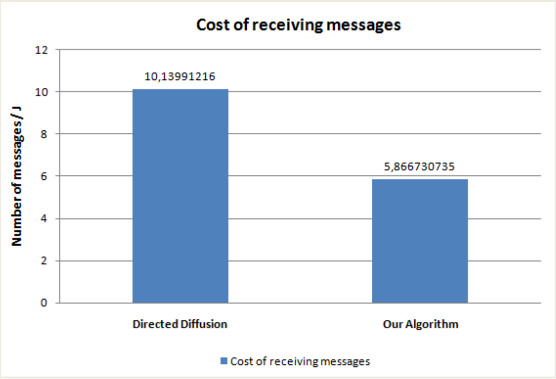


Figure 8. Cost of receiving messages

The Figure 8 shows that the cost of receiving messages is reduced by approximately 50 percent, using our proposal. This is due to the reduction in the number of control messages and the selection of routes that have a smaller number of hops with higher residual energy.

V. Conclusion

We presented, on this article, a routing algorithm that aims at reducing the energy consumption of the network. The proposal consists in the modification of the Directed Diffusion protocol. Two modifications were implemented in the original protocol: the first one is a geocast approach and the second one is a mechanism for selection of routes. The modifications have the purpose of reducing the flooding caused by the Directed Diffusion in order to recover broken paths and to determine the best route for delivery of the collected event.

The evaluation of the proposal, by means of simulation, demonstrated that the utilization of the geocast approach and of the mechanism for selection of routes shows positive results for reducing and balancing the energy consumption of the network.

It is intended to evaluate, in future works, the cost for receiving messages, which measures how many messages are delivered for each energy unit (J) used for all of the network nodes. In addition to that, we will also evaluate the reliability of the network, which will be measured based on the delivery rate.

Acknowledgment

This study was partially sponsored by CNPq (National Counsel of Technological and Scientific Development) process Nr. 135654/2008-4 and 112816/2009-6.

References

[1] Chen, B., Jamieson, K., Balakrishnan, H. and Morris, R. Span: An energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks. In *7th Annual International Conference on Mobile Computing and Networking* (July 2001).

[2] Shah, R. C. and Rabaey, J. Energy aware routing for low energy ad hoc sensor networks. *IEEE WCNC* (2002).

[3] Lee, Sung-Ju; Hsu, J.; Hayashida, R. Gerla, M. et al. Selecting a Routing Strategy for your Ad Hoc Network. University of California, Los Angeles. Computer Communications, Vol. 26, Issue 7, pp 723-733, May 2003, Elsevier Science.

[4] Y.-B. Ko and N. H. Vaidya. Location-aided routing (LAR) in mobile ad hoc networks. Wireless Networks, Hingham, MA, USA, v.6, n.4, p. 307–321, 2000.

[5] Y.-B. Ko and N. H. Vaidya. Location-aided routing (LAR) in mobile ad hoc networks. In *ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'98)*, October 1998.

[6] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed Diffusion: A scalable and robust communication paradigm for sensor networks", *Proceedings of ACM MobiCom'00*, Boston, MA, pp. 56-67, Aug. 2000.

[7] Y.-B. Ko and N. H. Vaidya. Optimizations for location-aided routing (LAR) in mobile ad hoc networks. Technical Report 98-023, CS Dept., Texas A&M University, November 1998.

[8] Y.-B. Ko and N. H. Vaidya. Location-based multicast in mobile ad hoc networks. Technical Report 98-018, CS Dept., Texas A&M University, September 1998.

[9] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirici, "A survey on sensor networks", Communications of the ACM, pp. 102-114, Aug. 2002. A.P.R. Canázio, Automatic Cork Processing System, Reference Manual. Addinson Wesley Publishing Company, 2002.

[10] Teixeira, I., de Rezende, J. F., and Pedroza, A.C.P. - "Wireless Sensor Network: Improving the Network Energy Consumption", in XXI Simpósio Brasileiro de Telecomunicações - SBT'2004, Belém, Brazil, September 2004.

[11] Sinalgo (2007). http://dcg.ethz.ch/projects/sinalgo/.

[12] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan,"Energy-Efficient Communication Protocol for Wireless Microsensor Networks". Hawaii International Conference on System Sciences, January 4-7, 2000, Maui, Hawaii.

Conflicts in Online Social Network Privacy Policy

Analysis and Resloution

Zhengping Wu
Department of Computer Science and Engineering, University of Bridgeport
221 University Avenue, Bridgeport, CT 06604, USA
zhengpiw@bridgeport.edu

Yuanyao Liu
Department of Computer Science and Engineering, University of Bridgeport
221 University Avenue, Bridgeport, CT 06604, USA
yuaoyaol@bridgeport.edu

***Abstract*—Social network sites have the potential to fundamentally change the character of our social lives, both on an interpersonal and a community level. In the past few years, social network sites have become integrated into the daily practices of millions of users. Social applications provide a convenient approach where users can communicate and share their personal information. However, privacy and information security have been concerned as a critical issue in development social applications. Policy-based management, which is a management approach, has been used to control the information access privilege. Users' policies are more flexible and ease to change. However, different policies may cause conflicts that affect execution of social applications. In this paper, we propose a semantic temporal logic mechanism to detect policy conflict in social network environment.**

Keywords-component;

I. INTRODUCTION

Social network sites have become increasingly popular; many people are using Facebook, Myspace, and Twitter to communicate with others. Social network sites are not a platform for communication; it has become a collaboration platform, since social applications provide a great extension. People share their information with others on social network sites, but the privacy control is simple and limited through a number of selections that social network sites provide.

The main motivation for the users to create accounts and participate in these services is sharing their information with others for some purposes. People usually sharing two types of information: professional information and private information. In the first case, social network sites are as a utility to establish and entertain business task. In the second case, social network sites are used to share private information such as personal contract details, personal photos and videos. However, by using social applications, social network sites are not just used to share information, but to communicate with specifically people. Sharing information has become information exchange.

OpenSocial is a set of common application programming interfaces (APIs) for web-based social network applications, developed by Google along with MySpace and a number of other social networks. Applications that build with OpenSocial APIs can be implemented into different social network sites, which support OpenSocial APIs. Social applications are information exchange utilities, which will acquire user's information from social network sites, where users put their information, and then provide some service base on this information. Social applications usually provide some privacy control settings, which will be enforce into social network site where the application is installed. It means if a same user installs one application on two different social network sites, each site will have one privacy control setting. If this application is used to publish user signed information, two different sites may publish different information, which belong to a same user. This is a problem.

Privacy concern has been a critical issue since social network sites appear. Privacy control that social network site provide is limited to usage of social network site self. The privacy control in social applications has to be set by users or follow the social network site's setting. Since online social privacy is harder to guarantee [1], privacy concern has gotten more attention from society. Policy-based management has been used in social network sites. The privacy settings in social network sites are a kind of policy rules. These policy rules are easy generated into machine-readable format. Policy-based management is an administrative approach to manage system behaviors and rules within an information domain. A policy describes the set of behaviors that a party is required or allowed to take. In social network environment, these behaviors are usually access actions. There are two main types of policies in social network sites, one is global policy, which is used to control entire information set within user account, and another type is application policies. Application policies are used to control information that will be used in application processing.

Policy conflicts can be classified into static conflicts and dynamic conflicts. Analysis of static policy conflicts is time-independent. These conflicts can be resolved by simple logic reasoning or compiling techniques. Dynamic policy analysis is time-dependent, which means the applicability of policies and information will change over time. The major issue about

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_100,

dynamic policy analysis is the changing information. The information covered by policies can be identified as different types of elements. Relations between different elements are critical to conflict detection and resolution.

There are two platform-independent privacy languages: IBM's Enterprise Privacy Authorization Language (EPAL) [2] and the OASIS eXtensible Access Control Markup Language (XACML) [3] along with *Privacy policy profile of XACML* [4]. XACML is a large, complex language, and this can make formal analysis more difficult. While EPAL and XACML are very similar, and share a common authorization model, they differ in important ways. In almost every area, the functionality of XACML 2.0 is a superset of EPAL 1.2 [5]. Where the two languages differ, the EPAL differences usually result in less functionality than XACML has. In many cases, the EPAL 1.2 differences from XACML make construction of flexible or scalable privacy policies impossible or difficult. However, there is not a specification for social applications. In this paper, we will use a simple XML format to present privacy policies.

II. RELATED WORK

Privacy within online social network sites is often not expected or is undefined [6]. Dwyer [1] gives us an image that different online social network sites make users have different privacy concern level. The social application, which is an add-on for social network site, also affects the reputation of social network sites. In another hand, the social application is installed into user's interface (social applications will retrieve user's identify information and set up a connection between user and application server if it is needed). The social application always has more privilege to access user's information, like Facebook's APIs give developers more rights to access user's friends list. In this case, social applications can get more information than user allowed. Therefore, privacy concern on social applications is more than online social network sites themselves.

Police-based systems are suitable solutions for automated online collaborations across domain boundaries, but policies become more and more complex for cross-domain activities, which require more adaptive management. Since a request or an action from a foreign domain may not have an exact match with an existing policy, several compatible policies can be applied to it. Choosing one of these policies may lead to a conflict with another related policy or violate the principle of least privilege [7]. Even a similar request or action with different parameters may lead to difficulties when applying a previously compatible policy. In general, whenever multiple policies apply to an action or request, there is a potential for some form of conflict, but it is essential that multiple policies should be applied in order to cover the diversity of management needs and of different groups of users. There may be different policies related to security, privacy, or other management requirements, which are applied to different aspects reflecting different management needs. So policy conflict analysis (detection and/or resolution) becomes an essential part in management of policy-based systems.

On online social network sites, users define policies; they may not have ability to make sure policies on different online social network sites are consistent. If there are positive and negative authorization or obligation policies applied on the same subjects, targets and actions, modality conflicts may arise [8]. Modality conflicts can be detected through static analysis of the policy specifications. Once all policy-driven actions of a system have been analyzed and these different types of conflicts that can arise have been identified, it is possible to define rules that can be used to recognize conflict situations in a policy set. Then these rules can be invoked during a conflict detection process prior to policy deployment and enforcement to identify potential inconsistencies. This is known as static conflict detection and takes place at specification time. Most existing work in detecting conflicts are related to modality conflicts. But sometimes application-specific conflicts cannot be detected directly from policy specifications without additional information. Conflicts may arise from the semantics of the policies as well. Thus, a metapolicy [9] concept is introduced to represent this additional information, which is a constraint or context associated with related policies. Yet there is still a possibility of conflicts even when there is not any shared subject, targets or actions between two policies. There may be implicit relationships between subjects, targets and actions when policies are applied to them. Thus a metapolicy-based approach cannot resolve all conflicts, because an administrator would have to specify metapolicies for all possible runtime conflict situations. It is almost impossible to predict all implicit relationships that may cause conflicts in the system by a user.

In classical logic, propositional logic is based on a set of elementary facts by using a set of logic operators. It indicates a Boolean value set. First order logic (FOL) [10] is an extension of propositional logic. Higher order logic [11] is an extension of first order logic. Temporal logic [12] is an extension of classical propositional logic, which is built for representing the set of domain elementary facts by using a set of logic operators [12]. And ontology [13] is an explicit specification of a conceptualization, which represents a set of entities and the relationships between these entities in a specific domain. Entities are abstracted from real world objects in that domain; relationships are contacts or effects between objects. The ontology describing this type of information uses a set of formal terms and operators.

III. POLICY MODEL

3.1 General Policy Model

As a policy can contain one or more rules, which describe relations between policy executants and requester (subject and object), we focus on a simple policy that contains one rule. For more complex policies, they can easily be divided into simple policy segments. A policy segment describes a behavior and attributes about this behavior. For each behavior, there is an executor or a type of executors, an object or a type of objects, and some constraints, which limit and describe the action. So in each policy, there are four major components: subject, object (resource), action and context constraints. A policy segment describes a complete behavior, which contains subject, object, action, and context constraints. These elements are consisted of attributes, where detail information is stored. The policy's subject is a set of attributes (A) identifying the action's executor, it can be denoted as s={A}; the object is a set of attributes (A) that describes an action's target, it can be denoted as o={A}; context constraints include other attributes within the policy, such as network contexts, execution environments, and etc. We distinguish this type of context information to other attributes that are about subjects and objects. A policy's action actually represents a temporary binding between subjects and objects, which can be denoted as A(c)=s×o. Following notations show the components of a policy segment in this general policy model:

A: attribute — the attributes that describe the subjects and objects.
S: {As} — the subject of a segment, the set of attributes that limit policy subject.
O: {As} — the object of a object, the set of attributes that limit policy objects.
A(c)=S×O — the action of segment, it indicates the relation between subject and object.
Context — the information that hide in the network or environment.

In this policy model, each element is denoted in a set of attributes. In another words, a policy is a set of attributes with different constrain conditions.

3.2 Privacy Control Policy Model

A privacy control policy should contain a resource (object), a requester (subject), an action and a set of information that limited these elements. The subject is the information requester, who wants to access the information; the object is the resource or the information; the action is usually access or other action that requester want to perform. Therefore, the privacy policy model can be presents as follow:

s = {As}	s: information requester.
o = {As}	o: the user's information.
A(c) = s×o	A(c): the action that requester want to perform.
Context: Additional information	Context: the additional information to describe s, o and action.
Rule = {s×o×A(c)×Context}	Rule: a functional segment of a policy

Table 1 Privacy Control Policy Model

A rule is a functional segment of a policy; it indicates a condition, which is used to filter requests. Here are two privacy control policies that are used to check requester's role information and time information. Here is an example scenario: an organization collaborates with a research group in a university. They use online social network site as the collaboration platform. The social application provides the information exchange functionality. Here are two policies about the update action of different participants.

Policy 1

```
<?xml version="1.0" encoding="utf-8"?>
<Policy>
  <Workspace>
    <Update>
      <Permit>
        <Group name="Colleague"/>
        <User name="Allen"/>
      </Permit>
      <Period start="08:00:00" end="13:00:00"/>
    </Update>
  </Workspace>
</Policy>
```

Policy 2

```
<?xml version="1.0" encoding="utf-8"?>
<Policy>
  <Workspace>
    <Update>
      <Permit>
        <Group name="University"/>
      </Permit>
      <Period start="12:00:00" end="17:00:00"/>
    </Update>
  </Workspace>
</Policy>
```

In policy 1, the subject has attributes, which is role information. It indicates the request from the "Colleague" group members and specific user "Allen" will be permitted within a time period. In policy 2, the subject has the same type attributes, which is also role information. It indicates the request from the "University" group members will be permitted within a different time period compared to the first policy. Actions in these two policies are the same. The action reflects the relation between subjects and objects. In this case, action "Update" is showing that resource will be changed by the subject. Obviously, there is a conflict between these two policies.

Between two policies, the dynamic conflict will involve any element within those policies. We generated some situations that can be used in conflict analysis:

a) Identity may be replaced by role information to act as the subject ;
b) The identity information is provisional;
c) The resource is provisional;
d) The relationships (action) between a resource and an identity are provisional;
e) The relationships (action) are constrained by network or security domain contexts

IV. SEMENTIC TEMPORAL LOGIC

4.1 Temporal Logic and Ontology

One major issue about dynamic policy analysis is that information fluent will change over time. The dynamic policy analysis has to keep track of the change of information fluent.

It is easy to represent static states and values using classical logic. However, time-dependent situations cannot be represented by classical logic. Temporal logic also called tense logic can handle these time-dependent situations. Temporal logic is an extension of classical propositional logic, which represents a set of domain elementary facts using a set of logic operators [14]. It has been broadly used to cover all approaches to the representation of temporal information within a logical framework [15].

In Policy 1& 2, time period is a very important attribute. This attribute affects the effect of these policies.

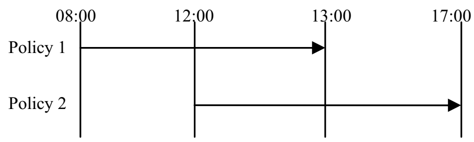


Figure 1 Policy 1&2 Time Period

The Figure 1 is time period of two policies. In traditional logic, it is hard to express this attribute. However, in temporal logic, we can present this situation as:

HoldsAt(colleague,Workspace,update(),t1);
Terminate(colleague, t3)
HoldsAt(University,Workspace,update(),t2);
Terminate(colleague, t4)
t1=08:00:00, t2=12:00:00, t3=13:00:00, t4=17:00:00.

The temporal logic can denote the state of elements and the changing of states. However, temporal logic does not contain information about relation with these policies. We need something that can express and store relation information about elements with a policy domain. The policy domain is a collection of elements and services, administered in a coordinated fashion [14].

Ontology defines a set of representations of classes, instances, attributes, and relationships among them. Therefore, the ontology-based knowledge representation is a good tool for the purpose of information transferring, sharing, and analysis. It is a good support for our proposed Semantic Temporal Logic (STL).

4.2 Semantic Temporal Logic (STL)

Using temporal logic to process cross-domain or multi-domain policies, semantic relationships among different components is unavoidable, which reduce the accuracy and efficiency of policy analysis. STL provides a bridge over this barrier by applying an ontology-based knowledge representation onto temporal logic itself. This representation provides not only information of individual entities from specific domains but also the relationships among these entities, which are key issues for policy analysis.

The ontology-based knowledge representation stores the information of specific policy domain, which contains each entity's information and relationships among several entities. Back to Policy 1 and Policy 2 described in the previous section, there are some attributes within these policies:

Attributes : Colleague.time={08:00,13:00};
Update.permit={true};
University.time={12:00,17:00};
Update.permit={true};

The relationship between "Colleague" and "University" is that they play mutually exclusive actions, and there are some restrictions.

Relation: not_Update (Colleague, University, t);

To build the representation foundation, we use ontology to describe the model of policy domain and relations within that policy domain. This ontology keeps track of the relations between elements in the policy domain. Ontology describes the elements in a domain and also the relationships between these elements. Different ontology languages provide different facilities. Using OWL-Full, the above policy example, "Colleague" is a class, which has some properties such as "Update" and "from 08:00 to 13:00". The "University" is another class, which has properties such as "Update" and "from 12:00 to 17:00". The relation between "Colleague" and "University" is "they play same action", and the property on this relation is mutually exclusive (12:00 to 13:00). And we can see there is another element in Policy 1, user "Allen" is also allowed to update the object during "Colleague" time, however, we do not know other attributes about this special element. Therefore, when we analyze these policies, we have to check properties of "Colleague", "University" and user "Allen":

Policy= {"Colleague", "Allen"}×"Workspace" ×"Update" ×Period;
Policy= "University" ×"Workspace" ×"Update" ×Period;

V. DYNAMIC POLICY ANALYSIS

5.1 Dynamic Policy Analysis

Dynamic policy conflict analysis is then based on logic reasoning and Boolean function comparison. The complexity of policy rules and the environment fluent always affect the result of analysis. And also, the policy conflict analysis needs lots of manual work to adjust semantic relationships for accuracy, especially distinguishing subjects and objects in policy segments. Temporal logic can be used to perform automatic policy conflict analysis for dynamic policies, but the accuracy cannot be guaranteed due to the lack of capability to support semantic relationships among entities. The proposed Semantic Temporal Logic (STL) uses ontology as the knowledge supplier, which improves accuracy and reduces ambiguity in analysis results. The ontology, which records the policy model and relationships within the policy model, helps the logical reasoning mechanism identify elements and relationships between elements. For example, Policy 1 and Policy 2 are two privacy control policies express the relations between two elements: Colleague and University.

Logical expressions for this example are:

HoldsAt(Colleague,Workspace,update(),t1);
HoldsAt(University,Workspace,udpdate(),t2);
08:00 < t1 <13:00;
12:00 < t2<17:00;
Relation: not_update(colleague, university, t3);
12:00<t3<13:00 ;

So the STL rules for this example become:

HoldsAt(colleague, Workspace,Update, t1)
Terminate(colleague, t2)
HoldsAt(University, Workspace, Update,t3)→
Conflict(Update(colleague, Workspace)&&Update(University, Workspace),t)

So in a rule of STL, semantic relationships are combined with logic predicates to represent policies from multiple domains with dynamic elements. This rule in the above example uses the relationships between different entities to detect dynamic conflicts among different policies.

Privacy-control policies often place requirements on the actions of more than one subject. The dynamic environment could affect the enforceability of policies. This dynamic environment could be expressed by ontology. The ontology contains relations, attributes and other restrictions. The information within this ontology supports the temporal logic evaluations. We use the privacy policy model discussed before as the foundation of dynamic policy analysis.

Here are STL rules that are used to detect conflicts; these rules are also used to return conflict type.

- *HoldsAt(permit(Role1(sub),obj,A1(c)),t) ∧ HoldsAt(permit(Role2(sub),obj,A2(c)),t) ∧ A1<>A2 → HoldsAt(overlapConflict(conflictOfDifsubject,overlaps(permit(Role1(sub),obj,A1(c)),permit(Role2(sub),obj,A2(c)),t) → Trajectory(permit(Role1(sub),obj,A1(c)),t,deny(Role2(sub),obj,A2(c)),d) ∨ Trajectory(permit(Role2(sub),obj,A2(c)),t,deny(Role1(sub),obj,A1(c)),d)*
  Relation: sub.attribute.role1 ≠ sub.attribute.role2;
  begin<t<finial;

[The situation a: the subject (sub) can have different two roles (Role1(sub) and Role2(sub)). When different roles perform different actions (A1(c) and A2(c)) toward the same object (obj) at time t, then an overlap conflict may occur. Only one action will be permitted.]

- *HoldsAt(doAction(sub1,obj,A1(c)),t) ∧ HoldsAt(doAction(sub2,obj,A2(c)),t) ∧ A1<>A2 → HoldsAt(dynamicConflict(conflictOfDifsubject,Overlaps(doAction(sub1,obj,A1(c)),doAction(sub2,obj,A2(c)),t),t) → Trajectory(permit(sub1,obj,A1(c)),t,deny(sub2,obj,A2(c)),d) ∨ Trajectory(permit(sub2,obj,A2(c)),t,deny(sub1,obj,A1(c)),d)*
  Relation: at time t,sub1.attribute.Id ≠ sub2.attribute.Id;
  HoldsAt(sub1,obj,A1(c),t);
  HoldsAt(sub2,obj,A2(c),t);
  begin<t<finalt

[The situation b: when two different subjects (sub1 and sub2) use same authentication secret (obj) perform different action toward the same target at the same time t, another type of overlap conflict may occur. Only one action will be performed on the object.]

- *HoldsAt(permit(Sub,Obj1,A1(c),t1)) ∧ HoldsAt(permit(Sub,Obj2,A2(c),t2)) ∧ t1<=t<=t2 ∧ Clipped(t1,authorize(sub,secret,t),t2) → HoldsAt(dynamicConflict(conflictofInterest,overlaps(permit(Sub,Obj1,A1(c),t1),permit(Sub,Obj2,A2(c),t2),Clipped(t1,authorize(sub,secret,t),t2)),t),t) → Trajectory(permit(Sub,Obj1,A1(c),t1),t,deny(Sub,Obj2,A2(c),t2),d) ∨ Trajectory(permit(Sub,Obj2,A2(c),t2),t,deny(Sub,Obj1,A1(c),t1),d)*
  Relation: HoldsAt(sub,obj1,A1(c),t1);
  change(obj1,obj2,t);
  t1<t<t2;

[The situation c: a subject (sub) may perform different actions (A1(c) and A2(c)) toward a changing object (authentication secret) in the time period between t1 and t2. A conflict of interest may occur. Then only one action will be permitted.]

VI. IMPLEMENTATION AND EXPERIMENTS

6.1 Privacy Policy Conflict Analysis

We design a prototype system that can dynamically analyze privacy policy to detect potential conflicts within an online social network collaboration environment. The system is implemented in Microsoft .NET platform. We use its subset OWL-Full to store and retrieve our ontology-based temporal logic STL.

This prototype is designed to analyze cross-domain privacy control policies and indicate whether there is any conflict when a cross-domain collaboration activity happens. The prototype contains two parts, one is temporal logic evaluation, which is the logic mechanism, and the other one is ontology model, which contains relations within policy model.

6.2. Experiments

The policy set used in the experiment contains 5 pairs of privacy control policies. Each pair of policies indicates one type of conflicts we mentioned before. And there are 45 random combinations, which are not all conflict. The number of conflicts of these 45 pairs is 21. The prototype analyzes these policy pairs based on different types of conflict, which we discussed before. If use temporal logic to analyze these policy pairs without ontology part, the accuracy will decrease. The Figure 2 illustrates a comparison of the accuracy when using STL and using general temporal logic analysis result. From this figure, we can see that STL could easily analyze both designed policy pairs and random policy pairs with higher accuracy than general temporal logic.

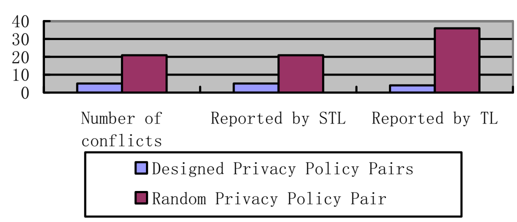


Figure 2 Comparison of the accuracy

VII. DISCUSSION

Temporal logic has been used to analyze required properties in trust policies, such as the FTPL mechanism [16], which use first order temporal logic to check satisfaction of states. In 2003, Event Calculus was proposed to analyze a combination of authorization and management policies [17]. But the authors do not identify the special capability of Event Calculus for transient properties in various activities. We apply Event Calculus in STL for analyzing transient properties and actions (relationships) in collaboration environment and find its capability to detect and resolve dynamic conflicts.

Meanwhile, static and dynamic conflict detection and resolution in policy-based management for large open distributed systems are discussed in [18]. A general model of security policies has been discussed in [19]. Detection and reconciliation of security policy conflicts following that model are restrained by the complexity of the policy set to be reconciled. And only two-party conflict reconciliation can be tractable. Table 2 illustrates these comparisons.

	Dynamicity	**Policy domain**	**Conflict detection and resolution**
STL	Dynamic and Static	Multiple	Detection and resolution suggestion
FTPL	Static	Single	Detection
Conflict detection in [18]	Static and Partial dynamicity	Single	Detection and resolution suggestion
Ismene in [19]	N/A	Multiple	Detection and resolution suggestion

Table 2 Comparison of Different Policy Analysis Mechanism

VIII. CONCLUSIONS

In this paper, we use Semantic Temporal Logic for analysis of privacy policy conflict in an online social network collaboration environment. Through the experiments on our prototype system, the improvement on capability and accuracy of Semantic Temporal Logic for dynamic privacy policy analysis is confirmed. A comparison of different policy analysis mechanisms shows that the ontology-based policy analysis system is more flexible, accurate, and applicable.

REFERENCES

[1] Catherine Dwyer, Starr Roxanne Hiltz, Katia Passerini , "Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace", Proceedings of the Thirteenth Americas Conference on Information Systems, Keystone, Colorado August 09 - 12 2007.

[2] IBM, Enterprise Privacy Authorization Language (EPAL), Version 1.2, 2003; http://www.w3.org/Submission/2003/SUBM-EPAL- 20031110/.

[3] Moses, T., ed., eXtensible Access Control Markup Language *(XACML), Version 2.0*; OASIS Standard, February 1, 2005; http://www.oasis- open.org/committees/ tc_home.php?wg_abbrev=xacml.

[4] Moses, T., ed., *Privacy policy profile of XACML v2.0*; OASIS Standard, February 1, 2005; http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-privacy_profile-spec-os.pdf.

[5] Anderson, A., Comparing Two Privacy Policy Languages: EPAL and XACML, Sun Microsystems Laboratories Technical Report 2005-147, 2005;
http://research.sun.com/techrep/2005/smli_tr-2005-147/TRCompareEPALandXACML.html.

[6] Dwyer, C. (2007) "Digital Relationships in the 'MySpace' Generation: Results From a Qualitative Study." Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS), Hawaii, 2007.

[7] Liang Chen, Jason Crampton, "Inter-domain role mapping and least privilege," *Proceedings of the 12th ACM symposium on Access control models and technologies*, pp. 157-162, 2007.

[8] M.Sloman, "Policy Specification for Programmable Networks," *Proceedings of the First International Working Conference on Active Networks*, pp: 73 – 84, 1999.

[9] M. Sloman, E. Lupu, "Security and Management Policy Specification," *IEEE Network*, vol. 16, no. 2, pp. 10-19, March/April 2002.

[10] Nir Friedman, Joseph Y. Halpern, Daphne Koller, "First-order conditional logic for default reasoning revisited," *ACM Transactions on Computational Logic (TOCL),* Vol. 1, Issue 2, PP. 175 – 207, 2000

[11] Andrews P.B., "An Introduction to Mathematical Logic and Type Theory: To Truth Through Proof, 2nd ed." *Kluwer Academic Publishers,* PP: 201-203, 2002

[12] P. Bellini, R. Mattolini, and P. Nesi, "Temporal logics for real-time system specification," *ACM Comput. Surv.*, vol. 32, no. 1, pp. 12-42, March 2000.

[13] B. Orgun, M. Dras, A. Nayak, G. James, "Approaches for semantic interoperability between domain ontologies," *Proceedings of the second Australasian workshop on Advances in ontologies,* Vol. 72, PP: 41-50, 2006

[14] A. Westerinen, J. Schnizlein, "RFC3198 - Terminology for Policy-Based Management," http://www.faqs.org/rfcs/rfc3198.html 2001

[15] Edward N. Zalta, "Temporal Logic", *the Stanford Encyclopedia of Philosophy*, fall 2008 Edition,
http://plato.stanford.edu/archives/fall2008/entries/logic-temporal/

[16] Arun K. Eamani, A. Prasad Sistla, "Language based policy analysis in a SPKI Trust Management System," *Journal of Computer Security*, Vol. 14, No. 4, pp. 327-357, 2006.

[17] A.K. Bandara, E.C. Lupu, and A. Russo, "Using event calculus to formalise policy specification and analysis," *Proceedings of IEEE 4th International Workshop on Policies for Distributed Systems and Networks*, pp. 26- 39, 2003.

[18] Nicole Dunlop, Jadwiga Indulska, Kerry Raymond, "Methods for Conflict Resolution in Policy-Based Management Systems," *Proceedings of 7th International Conference on Enterprise Distributed Object Computing*, pp. 98-109, 2003.

[19] Patrick McDaniel and Atul Prakash, "Methods and limitations of security policy reconciliation," *ACM Transactions on Information and System Security,* Vol. 9, No. 3, pp. 259-291, 2006.

Developing a Feedback Model for Managing Congestion in Packet Switched Wide Area Networks

By

*E.F. Ogbimi, A.O. Oluwatope, G.A. Aderounmu
Department of Computer Science and Engineering
Obafemi Awolowo University,
Ile-Ife, Osun State, Nigeria
voksy01@yahoo.com

ABSTRACT

The objectives of the paper was to design a feedback model for congestion control in packet switched WAN and carry out the performance evaluation of the model. A feedback approach was generated by designing a feedback model. Control theory was used to assess the stability of the **bandwidth. The stability of flow of bandwidth was determined by finding the transfer function of the routers. A fluid flow model was developed using a queuing model. Secondly, the flow of bandwidth was optimized and the stability was checked. The conditions for stability in a packet switched WAN were established.**

1 INTRODCTION

1.1 Background of Packet Switched WAN

The network evolution has been sparkled by the increased application of exponential improvements in fundamental software and hardware technologies. These technologies are enabling more distributed network architecture characterised by the adding intelligence to the core of WAN, integrated multiprotocol/multiservice transport capability and variety of new services and service capabilities. The rapid growth of WAN in recent years is part of the struggle for network administrators to monitor the network for demand resources such as voice, video and data. Most users continue to add more networking devices of higher capacities to satisfy the needs of network users. Yet, the network remains congested and application software continue to run slowly and experience poor performance of networks. New breed of software application and information technology drivers like academic, industrialists and business tycoons contribute to this problem by installing new software which may not be favourable to the networking environment [9]. Academics, Industrialists and business tycoons demand for bandwidth at the same time which leads to the flow of bandwidth instability. The bandwidth demand for resources is higher than the capacity of the network in this case. In view of the various efforts of the network providers to sustain network performance, congestion and poor performance are still common features of networks. Hence, there is need to continue to conduct research on how to improve performance of packet switched WAN.

2 LITERATURE REVIEW

2.1 Related Research Works

Janarthanan, et al. (2003) worked on Enhancement of Feedback congestion Control by deploying active congestion control. They used a scheme called active congestion control to make the feedback congestion control more responsive to network congestion, using the active network technology. The work was done by designing the finite state machine models of every single component in the modules. The model was decomposed into set of states and working of the model embedded in the transition between states.

General AIMD (Addition Increase Multiplicative Decrease) congestion control was written by [25]. They proposed a window adjustment strategy where the window increases and decreases value used as parameters depending on the ACKS (Acknowledgements) received or the packet lost. The approach used was GAIMD (General Addition Increase Multiplicative Decrease). The analytic expression for sending rate of GAIMD was derived as

$$T_{\alpha,\beta}(P, RTT, T, b) = \frac{1}{RTT\sqrt{\frac{2b(1-\beta)p}{\alpha(1+\beta)}} + \left(\sqrt{\frac{(1-\beta^2)b}{2p}}\right)p\,(1+32p^2)} \quad (1)$$

Mahajan and Floyd (2001), wrote on RED-PD (Random Early Detection with Preferential Dropping). They used packet drop at the router to the preferentially drop packets of high-bandwidth flow to address congestion. The approach used was a partial state flow. In the approach, it identified high bandwidth during times of congestion and designed an architecture for RED-PD. Preferential dropping was done by computing probabilistically dropped packets from monitored flows before entering the output queue.

$$P\,(dp) = 1 - \frac{Target\,(R, \rho)}{Arrival\ rate} \quad (2)$$

ρ = drop rate, R = Target Response function

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_101,

2.2 A new feedback scheme

The new scheme being proposed was called dynamic explicit feedback model which will keep the bandwidth stable. It will reduce packet drops to its barest minimum. It will also reduce bandwidth delay in packet switched WAN. It will balance the flow of transmission through paths. A feedback mechanism is added to monitor the acknowledgement of packets from receiver to sender. It will also make bandwidth utilisation high, converge to fairness and minimize amplitude of oscillations and maintain high responsiveness.

3 METHODOLOGY

A feedback approach was generated by designing a feedback model for congestion control in packet switch WAN. The principle adopted was control theory. Control theory was used in this case to assess the stability of flow of bandwidth in a packet switched wide area network, when congested and when not congested. The stability of the bandwidth was determined by finding the transfer function of the flow of bandwidth of the network. The transfer function is done by finding the feedback gain of the routers in the network. In this paper, three major things were done. Firstly, a fluid flow model for packet switched WAN was developed using queuing model. Secondly the flow of bandwidth will be optimized and the stability was checked. Thirdly, the conditions for stability in a packet switched WAN were established. The result was validated using MATLAB. The model was implemented in MATLAB.

3.1 Control Theory

Control theory is defined as the mathematical study of how to manipulate the parameters affecting the behaviour of a system to produce the desired or optimal outcome [21].

The main conceptions of control theory are dynamic systems, desired outcome, its state and feedback. Control theory was concentrated mostly on the problems of stabilization.

The stability of the bandwidth was evaluated using feedback control theory. This was done by finding the transfer function of the arrival rate of packets. Transfer function is defined as the ratio of the output of the system to the input of the system [17].

Mathematically, Transfer function is defined as

$$H(s) = \frac{Y(s)}{X(s)} \quad (3)$$

Where *Y(s) = Output function of the system in Laplace domain*

X(s) = Output function into the system in Laplace domain

3.2. Model Construction

The figure below shows the block diagram which illustrates what this study was set out to achieve. The diagram shows transmission of packets from source (s) to destination using a feedback model. A message is transmitted to the first router which acts as a buffer. The message remains in the buffer for a while until there is an available router to send packets until the packets get to their destinations. The message is broken into packets by means of packetization. The packets are then distributed to the second and then third routers. There could also be many routers depending on the topology of the network and the environment the network is used. An acknowledgement will indicate that the packet has been sent to the receiver. Some of the packets could be lost, delayed. The feedback controller passes a feedback message to any of the routers that there is congestion somewhere along the link, then the router adjusts the speed of transmitting packets. The stability of flow of packets will then be regulated.

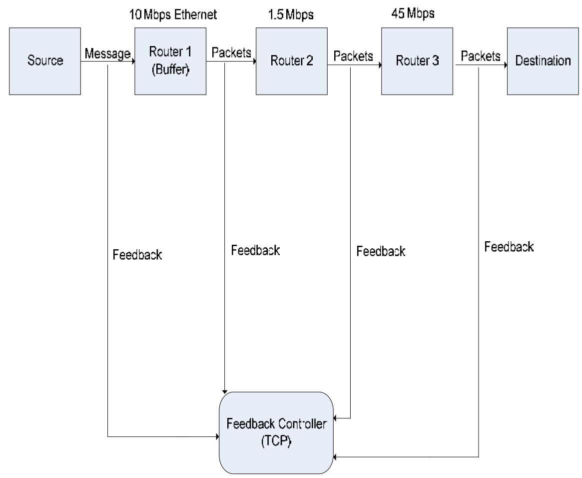


Figure 1: Block Diagram for Feedback model for managing congestion in Packet Switched WAN

3.3 Queuing Theory

Queuing theory is the mathematical study of waiting lines. Queuing theory is applicable in packet switched WAN because queues can be chained to form queuing networks. The use of queuing in packet switched WAN allows the users to wait for their request from any of the web servers to be completed and idle routers will be available for packets to be distributed.

3.4 Notation

Kendall's notation was adopted in this paper. It is notated as A|B|C; it denoted a queuing system having A as interarrival time distribution, B as service time distribution and C as number of servers or routers [20].

A convenient notation for summarizing the characteristics of queuing situations in a system is given by the following format: (K|M|C): (GD|∞|∞) [19], where:

K = arrival distribution
M = Service distribution
C = Number of servers in the network
GD = General discipline (all types of discipline)
First ∞ = Indefinite number of network users accessing the network
Second ∞ = Capacity of resources accessed or downloaded from the web

The steady-state measures of performance in a queuing situation are:.

λ = rate of arrival of packets
μ = service rate (rate of transmitting packets)
L_s = Average number of users on the network

L_q = Average number of users requesting for resources from a router in the network (in the queue).
W_q = Expected time users spend in requesting for resources from a router on the network.
W_s = Expected time users spend on accessing the network
P_n = Probability of routers that do not accept packets
$\frac{1}{\mu}$ = service time (time taken to transmit packets)
$\bar{c}$ = Expected number of routers that do not accept packets
c = number of routers ρ = network utilization

The model was applied to the situation of packet switched WAN. The parameters listed above were applied to the situation of packet switched WAN accordingly. The queuing model for packet flow in packet switched WAN was modelled by the following parameters above.
The difference between average number of users on the network and average number of people requesting for resources from the web server is denoted by this equation

$$L_s - L_q = \boldsymbol{\lambda}_{eff}/\mu \quad (4)$$

Network utilization = $\bar{c}$/c [19]
For a single server model (c =1), there is no limit on maximum number in the network. The model assumes a capacity source. Arrivals occur at rate **λ** user per unit time. ρ < 1 and **λ** < μ. If arrival rate is higher than the service rate, then the geometric series will not converge and the steady-state probability will not exist. The queue length will continually increase and no steady state will be reached in the network.
For multiple server models which is the area of concentration. There are c parallel servers (routers). The arrival rate is **λ** and the service rate per server is μ. There is no limit on the number of routers in the network. With Kendall's notation (K|M|C) : (GD|∞|∞)
K is the arrival distribution, M is the service distribution and C is the number of servers in the network. The queue discipline is generalized distribution that satisfies all conditions. The arrival rate and number of servers will be optimized in this case.
If $\rho = \lambda/\mu$ and assuming $\rho/c < 1$, the value determined from

$$\sum_{n=0}^{\infty} P_n = 1 \quad (5)$$

L_q is determined in MIMO (Multiple Input Multiple Output) systems

$$\text{If } \rho = \lambda/\mu \qquad (n = 0, 1, 2\ldots.., n)$$

Then $P_n = \dfrac{\lambda^n}{\mu \times 2\mu \times 3\mu \times \text{----------} \times n\mu}$

$$P_n = \frac{\lambda^n}{n!\mu^n} * P_0 \qquad n < c$$

$$= \frac{\lambda^n}{c!c^{n-c}\mu^n} * P_o \qquad n \geq c$$

$$L_q = \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} * P_o \quad (6)$$

$$L_s = L_q + \rho \quad (7)$$

$$L_s = P_0 \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} + \rho$$

$$= P_0 \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} + \rho\,(c-1)!\,(c-\rho)^2 \quad (8)$$

These are the measurements of performances in a multiple server and multiple queue in a packet switched wide area networks.
The number of network users requesting for resources from the web and network users on the network will be differentiated with respect to the arrival rate of packets. L_s and L_q will be differentiated with respect to **λ**.

$$L_q = P_0 \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} \quad (9)$$

Assuming c =1 for a single server to many number of users

$$L_q = P_0 \frac{\rho^2}{(1-2\rho+\rho^2)} \quad (10)$$

Network utilization = $\rho = \frac{\lambda}{\mu}$; insert $\rho = \lambda/\mu$ in equation (10)

$$L_q = P_0 \frac{\rho^2}{1-2\rho+\rho^2}$$

$$= P_o \frac{\left(\frac{\lambda}{\mu}\right)^2}{1-\frac{2\lambda}{\mu}+\left(\frac{\lambda}{\mu}\right)^2}$$

$$= P_o \frac{\frac{\lambda^2}{\mu^2}}{\frac{\mu^2-2\mu\lambda+\lambda^2}{\mu^2}}$$

$$L_q = P_o \frac{\lambda^2}{\mu^2-2\mu\lambda+\lambda^2} \quad (11)$$

Computation of transfer function

The number of users requesting for network resources from routers and number of users on the network (L_q and L_s) will be differentiated with respect to arrival rate of packets (**λ**). The differentiated equations will be linearized and differentiated. The Laplace's transform for the differentiated linear functions will be found and the transfer functions will be found based on the conditions and principles for finding transfer functions.

Properties of transfer function

A transfer function has the following properties
- The transfer function is defined for a linear time invariant system.
- The transfer function between a pair of input and output variables is the ratio of Laplace transforms of the differentiated functions of the output to the input.
- All initial conditions of the system are set to zero.
- The transfer function is independent of the input of the system [17]

The input and the output functions in both equations are not linear with respect to arrival rate (**λ**) and could not be used to compute the transfer function.
However, one of the prominent processes of stability analysis in non linear systems is the linearization of the equation. A non linear function could be made a linear one through the process linearization. Linearization is the method of

accessing the local stability of an equilibrium point of a system of non linear differential equations or discrete dynamical systems [10]. The typical equation for a linear function is

$$y = f(\lambda) + f'(\lambda)(x - \lambda)$$

where y could be interpreted as the output of the system and x the input of the system.

Quotient rule of differentiation will be used to differentiate both the number of users in both the network and requesting for web resources from the server with respect to arrival rate of packets.

From: $$L_q = P_o \frac{\lambda^2}{\mu^2 - 2\mu\lambda + \lambda^2}$$

Using Quotient rule of differentiation with respect to arrival rate of packets

$$\frac{dL_q}{d\lambda} = \frac{2P_0\lambda\mu}{(\mu - \lambda)^3} \qquad (12)$$

$$y = f(\lambda) + f'(\lambda)(x - \lambda)$$

at equilibrium point, $f(\lambda) = 0$

If y = 0, then

$$f'(\lambda)(x - \lambda) = 0$$

$$\frac{2P_0\lambda\mu\,(x - \lambda)}{(\mu - \lambda)^3} = 0$$

$$2P_0\lambda\mu\,(x - \lambda) = 0$$

$$x = \lambda$$

$$y = f(x) + f'(\lambda)(x - \lambda)$$

$$f(x) = 0$$

$$y = \frac{2P_0x\mu(x - x_0)}{(\mu - x)^3}$$

$$y = 2P_0\lambda\mu^2 + \mu^3 - 3\mu^2\lambda + 3\mu\lambda^2 - \lambda^3\,(x - \lambda)$$

If y = 0, then

$$x\left(\frac{2P_0\lambda\mu^2}{(\mu - \lambda)^3} + 1\right) = \lambda\left(\frac{2P_0\lambda\mu^2}{(\mu - \lambda)^3} + 1\right)$$

$$x = \lambda$$

$$y = f(x) + f'(x)(x - x_0)$$

$$y = \frac{2P_0x\mu^2 + \mu^3 - 3\mu^2x + 3\mu x^2 - x^3(x - x_0)}{(\mu - \lambda)^3} \qquad (15)$$

if $x_0 = 0$, then

$$y = \frac{2P_0x^2\mu^2 + \mu^3x - 3\mu^2x^2 + 3\mu x^3 - x^4}{(\mu - \lambda)^3} \qquad (16)$$

$$Y(s) = \frac{2P_0s^2\mu^2 + \mu^3s - 3\mu^2s^2 + 3\mu s^3 - s^4}{(\mu - s)^3}$$

$$sY(s) = \frac{2P_0s^3\mu^2 + \mu^3s^2 - 3\mu^2s^3 + 3\mu s^4 - s^5}{s(\mu - s)^3}$$

If $x_o = 0$

$$\frac{\mathcal{L}dL_q}{d\lambda} = sY(s) - Y(0)$$

$$\frac{dL_q}{d\lambda} - \frac{2P_0\mu s^2}{(\mu - s)^3}$$

For the number of users in the network:

$$L_s = P_0\frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} + \rho\,(c - 1)!\,(c - \rho)^2$$

For a single server model to many users
Assuming c = 1

$$L_s = \frac{P_0\rho^2 + \rho\,(1 - 2\rho + \rho^2)}{(1 - 2\rho + \rho^2)}$$

$$L_s = \frac{P_0\left(\frac{\lambda}{\mu}\right)^2 + \frac{\lambda}{\mu}\left(1 - 2\left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)^2\right)}{\left(1 - 2\left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)^2\right)}$$

$$= \frac{P_0\lambda^2\mu + \lambda\mu^2 - 2\lambda^2\mu + \lambda^3}{\mu^2 - 2\mu\lambda + \lambda^2} \qquad (13)$$

$U = \lambda^2\mu + P_0\lambda^2\mu - 2\lambda^2\mu + \lambda^3$; $V = \mu^2 + 2\lambda\mu + \lambda^2$

Using quotient rule of differentiation with respect to arrival rate of packets.

$$\frac{dL_s}{d\lambda} = \frac{2P_0\lambda\mu^2}{(\mu - \lambda)^3} + 1 \qquad (14)$$

$$y = f(\lambda) + f'(\lambda)(x - \lambda)$$

at the equilibrium point $f(\lambda) = 0$

$$y = \frac{2P_0\lambda\mu^2}{(\mu - \lambda)^3} + 1$$

$$\mathcal{L}\frac{dL_s}{d\lambda} = SY(s) - Y(0$$

$$Y(s) = \frac{2P_0s^2\mu^2 + \mu^3s - 3\mu^2s^2 + 3\mu s^3 - s^4}{(\mu - s)^3} \qquad (17)$$

Transfer function =

$$\frac{\mathcal{L}L_q}{\mathcal{L}L_s} = \frac{2P_0\mu s^2}{2P_0\mu s^2 + \mu^3s - 3\mu^2s^2 + 3\mu s^3 - s^4}$$

$$\frac{\mathcal{L}L_q}{\mathcal{L}L_s} = \frac{2P_0\mu s^2}{2P_0\mu s^2 + (\mu - s)^3} \qquad (18)$$

4 RESULTS AND DISCUSSION

4.1 Bode Plot

A bode plot is a graph of the logarithmic of the transfer function of a linear, time invariant system versus frequency, plotted with a log-frequency axis, to show the system's frequency response. It is a combination of bode magnitude plot (expressed in dB of gain) and a bode phase plot (the phase is the imaginary part of the complex logarithm of the complex transfer function. The magnitude axis of the Bode plot is usually expressed as decibels of power that is by 20 log rule: 20 times the common logarithm of the amplitude gain.

A bode phase plot is a graph versus frequency, also plotted on a log-frequency axis, usually used in conjunction with the magnitude plot, to evaluate how much a frequency will be phase-shifted , for example a signal described by: A sin (wt) may be attenuated but also phase-shifted.

4.2 Impulse Response

In signal processing impulse response or impulse response function of a dynamic system is its output when presented with a brief input signal called an impulse. An impulse response refers to the reaction of any dynamic system in response to external changes. In both cases impulse response describes the reaction of the system as a function of time (or possibly as a function of some other independent variables that parameterizes the dynamic behaviour of the system. In this case, packet switched wide area networks is a dynamic system, the external system may be messages and the output may be information produced from the network. In this case, the impulse response refers to the change in information over time in response to the initial message.

In control theory, impulse response is the response of a system to a Dirac delta input. This is also useful in the analysis of dynamic systems. The Laplace transform of the delta function is 1, so the impulse response is equivalent to the Inverse Laplace transform of the system's transfer function

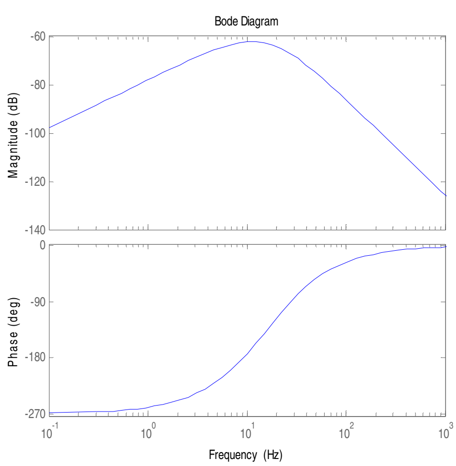


Figure 4.1.1: Bode plot of transfer function (h1) with a service rate of 100 and probability of number of users at 0.1.

$$h_1 = \frac{-20\,s}{S^3 - 300\,s^2 + 29980\,s - 1000000}$$

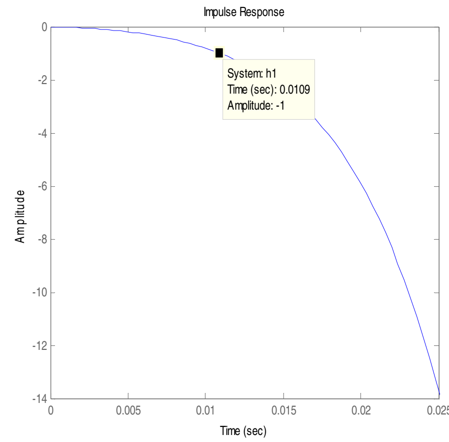


Figure 4.2.1: Impulse response of transfer function (h1) with a service rate of 100 and probability of the number of users is 0.1.

From figure 4.1.1, the axes moved up with a magnitude of -98dB and a frequency of 0.1 Hz until it got to a magnitude of -62.3dB with a frequency of 10.1Hz. The frequency of the bandwidth was stable at 10.1 Hz until it got to a frequency of 12.4 Hz. For the phase angle, the axis moved down with a frequency of 977 Hz and a phase angle of 0° and moved down till it got to a phase angle of -267° at a frequency of 0.322 Hz all through till it got to a frequency of 0.105 Hz. From figure 4.2.1, the Impulse response is a dirac delta which represents the limiting case of a pulse made in a very short time while giving a high peak. In this case, when the Amplitude is -1, the time taken to react to an external change is 0.0109 seconds.

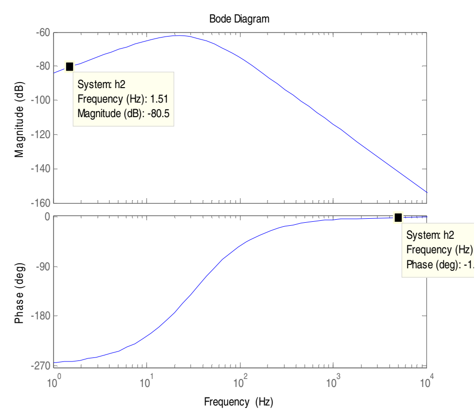
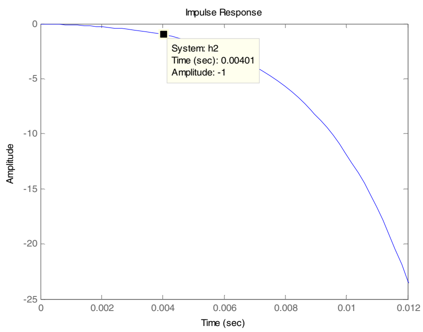


Figure 4.1.2: Bode plot of the transfer function with a service rate of 200 and probability of 0.1.

$$h_2 = \frac{-80\,s}{s^3 - 600\,s^2 + 119920\,s - 8000000}$$

Figure 4.2.2: The impulse response of the transfer function with a service rate of 200 and probability of number of users in the network was 0.2



From figure 4.1.2, the axes sloped up at a frequency of 1 Hz at -84 dB and then moved up until it became stable at a magnitude of -62.3dB with a frequency of 20.1 Hz and was stable till the axes got to the frequency of 24.8 Hz. The axes moved down and touched the frequency axis of 8000 Hz. Bandwidth became stable at the phase angle of -264°. The frequency which the bandwidth was stable was between 1.02 Hz and 1.19Hz. From the figure 4.2.2, the impulse response is a Dirac delta which represents the limiting case of a pulse made in a very short time given high peak. When amplitude is -1 then the time taken for the system to react to external changes is 0.004 seconds.

The magnitude of the bandwidth flow was -62.3 dB with frequency responses varying depending on the outcome of the arrival rate, service rate and probability of number of users in the network. The phase angle of the bandwidth flow stability was -267°. The flow of bandwidth could be stable with this condition and congestion could be reduced by using this model. The impulse function also shows that the packet switched wide area networks can exhibit an impulse within a small period of time in seconds depending on the optimization of arrival rates, service rates and probability of number of users accessing the network.

5 CONCLUSION

The Conclusion of this work is that congestion can be reduced by using a feedback model which is dynamic explicit feedback model. This was done by writing a queuing mathematical model for packet switched WAN. The bandwidth flow stability was measured by finding the transfer functions of the average rate of network users in the network and accessing the network. This was done by differentiating the number of network users and the number of network users requesting for resources from the web with respect to average rates of both network users accessing the network and registered users in the network. Their differentiated functions were linearised to make it a linear function. The laplace transform were computed and transfer function were also computed. The transfer function was then found and service rate, arrival rate, and probability of number of users were varied. The bode plots and impulse responses were drawn with MATLAB 7.0.

REFERENCES

[1] Borisenok, S.V., (2005), *Modern Control Theory and Its Application;* Department of Physics, Herzen State Pedagogical University, Moika River Embankment, 48, 191186, St Petersburg, Russia.

[2] Bosco, H.L. and D.C. Dowden, (2000), *Evolution of Wide Area Networks, Bellab Technical Journals.*

[3] Bose, S.J., (2002), Chapter 1 – *An Introduction to Queuing systems:* Kluwer/Plenum Publishers.

[4] Daintith, J. (1992), *Dictionary of Mathematics Series;* International Book Production, London.

[5] Dorf, R. C. and R.H. Bishop., (2004), *Modern Control Systems;* Pearson Education, Singapore.

[6] Flood, J.E., (1998) Telecommunications Switching, Traffic and Networks, Chapter 4: Telecommunications Traffic, New York: Prentice Hall.

[7] Ghosh, S.,(2004), *Control Systems: Theory and Applications;* Pearson Education, Singapore.

[8] Hellstein, J. and S. Parekh., (2008), *An Introduction to Control Theory with Applications to Computer Science,* IBM; T.J. Watson Research Centre.

[9] Hicks, M., (2004), *Managing Gigabit Networks, Applications and Services – Compuware;* Questnet 2204 Conference Networking Far and Wide Held at Cairns International Hotel, Cairns, Austrailia.

[10] http://en.wikipedia.org/linearisation.

[11] Jain, R., (1998), *Congestion Management in High Speed Networks*, Issues and Trends, IEEE Magazine, pp 24-30.

[12] Jain, R., (2006), *Congestion Management in High Speed Networks*, Viewed from (http://www.cse.ohio.state.edu/~jain/cis788-95/fp/atm_cong/index.htm) accessed 13th February, 2006.

[13] Janarthanan, Y., G. Minden, and J. Evans (2003), *Enhancement of Feedback Congestion Control Mechanisms by Deploying Active Congestion Control;* The University of Kansas Centre for Research, Inc., 2235 Irving Hill Road, Lawrence, KS 66044-7612. U.S.A.

[14] Mahajan R., and S. Floyd, (2001), *RED-with Preferential Dropping (RED-PD);* University of Washington , AT&T Centre for Internet Research at ICSI (ACIRI).

[15] Patterson, D.A., (2001), *Network*, Spring Lecture Slides; Department of Computer Science, University of California, Berkeley, U.S.A.

[16] Penittinen A, (2003), *Queuing systems*, Lecture Notes: S-38. 145-*Introduction to Teletraffic Theory*.

[17] Philips, C. L. And R. D. Harbor, (2000), *Feedback Control*, *Control Systems*, Fourth Edition, Prentice Hall, Upper Saddle River, New Jersey, U.S.A

[18] Shor, M.H., Li Kung, J. Walpole, D.C., Steere, Pul Calton., (2003), *Application of Control Theory to Modelling and Analysis of Computer Systems*; Department of Electrical and Computer Engineering, Oregon State University, Corvallis, U.S.A.

[19] Taha, H.A., (2008), *Operation Research: An introduction*, Eight Edition, Upper Saddle River, New Jersey, U.S.A. page 538

[20] Tijms, H.C.,(2003), *Algorithmic Analysis of Queues,* Chapter 9 in *A First Course in Stochastic Models, Wiley,* Chichester.

[21] Weisstein, E.W., (2003), *CRC Concise Encyclopaedia of Mathematics*; Boca Raton, London, New York: Chapman & Hall.

[22] Williamson C., (2001), *Congestion Control and Traffic Management in High Speed Networks;* Department of Computer Science, University of Calgary, 2500 University drive, New Calgary, AB, Canada T2N, 1N-1.

[23] www.nationmaster.com/queuing_theory

[24] www.sosmath.com/Linearization_Technique

[25] Yang, R.X., S.L. Lam., (2000), *General AIMD Congestion Control*; Department of Computer Sciences, The University of Texas, Austin, Austin, U.S.A.

[26] Zabczyk, J., (1992), *Mathematical Control Theory:* An Introduction; Springer Series, Institute of Mathematics, Polish Academy of Science, Sniadeckich, office 512, 00-956 Warsaw, Poland.

Voice Gateway to IP-PSTN Integration and NGN Architecture

Geneflides Laureno da Silva, Raimir Holanda Filho

Master's Course in Applied Computer Sciences – University of Fortaleza (Unifor)
Av. Washington Soares, 1321 – 60811809 – Fortaleza – CE – Brazil
geneflides@unifor.br, raimir@unifor.br

***Abstract*- Over the last decade, the technological environment has traveled paths of updates and consolidated the emergence of proposals for new architectures, aimed at integrating the various networks in the world of telecommunications and building synergy between data, voice and image. The NGN networks are based on the exchange of data packets, which permit a flexible and scalable environment where the key factor is the provisioning of new applications, enabling widespread mobility and integration of new services with legacy networks.**

To face with this reality, we proposed a solution which is adherent to the functional architecture of NGNs. According to the layered model, at the application layer, we developed five applications that permit the implementation of new services; at the control layer we use the open ip telephony platform asterisk and finally at the transport/access layer, our solution support the following protocols: SIP, IAX, ISDN, R2Digital and SS7 Over IP.

In order to validate our proposed solution, we implemented a Voice Gateway that provides an integration between IP-PSTN networks.

I. Introduction

Over the last decade, the technological environment has traveled paths of updates and consolidated the emergence of proposals for new architectures [1] [2], aimed at integrating the various networks in the world of telecommunications and building synergy between data, voice and image.

Telecommunications operators have made progress in modernizing it's network elements, trying to keep existing equipments and legacy networks and integrating this context with the new paradigm of next generation networks.

Small and medium-sized enterprises end up getting on the margins of these new structures due to the high cost of equipment required for this deployment, as well as the value of service provided by operators is high because the need for provision of dedicated access so much to voice as for data. When physical access does not become the point of difficulty involved, the price for adding new services is an important factor to be increased by the entry of small enterprises in this new technological order.

The industry has developed solutions for connection between IP and PSTN, but the vast majority are proprietary solutions and almost entirely be based on the binomial SIP-ISDN. We can mention [3] [4] [5] as proposals for developing services, standards and integration. In the industrial field, we can mention the work [6] as reference produced by a company that has contributed to the development of gateways aimed at integrating the IP world with the traditional telephone networks as well as the insertion of such devices to the new paradigm of networks, the IMS (IP Multimedia Subsystem). Such implementations have proven the existence of recurrent challenges that need to be addressed, especially if the focus moves towards the core of the majors telecommunications operators and consequently the provision of new services to their customers. One of the major challenges is being able to keep parity between the existing legacy networks and integrate new services from the new networks with ease and speed of development for these structures.

To face with this reality, we proposed a solution which is adherent to the functional architecture of NGNs. According to the layered model, at the application layer, we developed five applications that permit the implementation of new services; at the control layer we use the open ip telephony platform asterisk and finally at the transport/access layer, our solution support the following protocols: SIP, IAX, ISDN, R2Digital and SS7 Over IP.

II. NGN Networks

The NGN networks are based on the exchange of data packets, capable of providing telecommunications services making use of multiple bandwidths [7]. Through technologies that incorporate parameters of quality of service, the functions of service delivery and transport are treated independently. The key factors sought by next generation networks is that the provision of new services, enabling widespread mobility and integration of services with legacy networks.

The work[8] is an example which we can see the different aspects and workforce for NGN. We focused our solution in the ITU-T standardization.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_102, © Springer Science+Business Media B.V. 2010

The ITU-T (International Telecommunication Union - Telecommunications) is the international body that initiated studies on standardization of next generation networks. These activities began in 2003 and strengthened in 2004 with the formation of FGNGN (Focus Group New Generation Network) [9]. Is important to stress here, studies and referrals decided, the break occurring at seven specific areas of study, namely: capacity, architecture and functional requirements, quality of service, control, security, migration of existing networks to NGN and requirements for future packet-based networks. The definition of the basic architecture, interface specifications and implementation guidelines make up the document that became the ITU-T recommendation for NGN[10].

According to the work [11], we elaborate the Fig. 1 below, that represents a summary of the functional architecture of NGN and a basic description of the functionality performed by each layer.

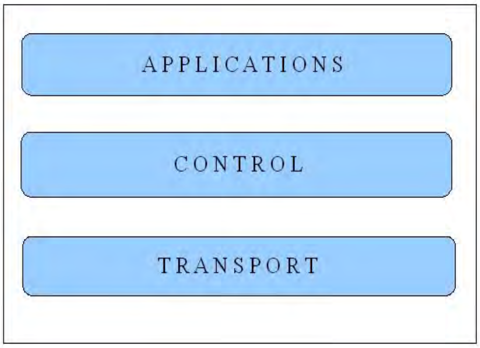


Fig. 1 NGN Functional Architecture

On the layered model used by the architecture of next generation networks, the transport layer has the responsibility to perform all connections between existing different networks, protocols and interfaces.

The provisioning of these various access networks, management and control of information are functions that are shared directly with the control layer. The IP protocol becomes the core technology to the world of next generation networks, since most of the equipment used, regardless of its functionality in the layered structure, uses this pattern to exchange signaling.

The control layer is the link between the transport layer and the features related to services and applications developed. This layer do the sharing of control functions of the transport layer, and control functions of the application layer. Availability of network resources, call setup, user terminals registration, accounting and authorization are examples of functions performed by the control layer.

The operation of applications like Charging and Billing are managed and controlled in this layer and more details about this kind of control can be found on the recommendation M3060 of ITU-T.

The main focus of application layer is to adding value to architecture, through the possibility of the provision of new services to existing IP networks and legacy networks.

The work [12] is an excellent example of consolidation and standardization of the services, permitting a better use of the integration of the telecommunication environments and patterns of development of the sphere of information technology.

The development of new applications quickly and following standards of development established, allow the creation of new services by operators in a way competitive and modular. Following this standardization, the application layer concentrates modules that can be connected or disconnected in a transparent manner.

In our view, the architecture consolidates its strong advantage and motivation in two main points: convergence between IP networks and legacy networks, creation and deployment of new services in way transparent and modular.

In the next session we describe our solution, that intend to support these two main points of NGN's contributions.

III. OUR SOLUTION

In our solution, the Gateway is fully implemented based on open libraries and open platforms. The Fig. 2 below, shows the adherence referred to the functional architecture of NGNs, summarizing the resources used and the applications developed.

For the application layer, five modules were developed with focus on generating services to users of the Gateway.

In the control layer, we use the open source platform Asterisk, whose main objective was to perform the interfacing between the applications developed and their access networks.

In the transport layer, we use standards already set and open libraries for providing interconnection with the following protocols: ISDN, R2Digital, SIP, IAX2 and SS7 over IP.

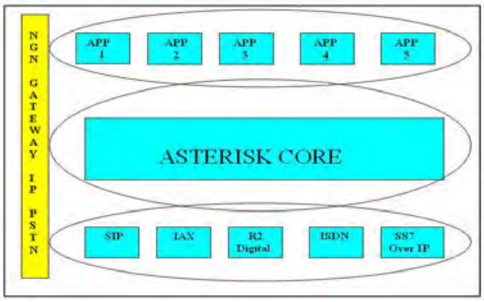


Fig. 2 A NGN Voice Gateway IP-PSTN Integration

We mention the work [13] as an example to test the use of multiple manufacturers and we showed that in our solution the validation of several protocols used have proven the flexibility and integration point between the IP-PSTN networks.

We developed five applications but here we will describe only two applications, the libraries used on the transport layer and the topology of the tests.

The five applications developed are as follows: Agent Authorization, Automatic Dialer, Charging, Callback and Employee Control. We will describe Automatic Dialer and Callback applications.

The first application that we will describe is Automatic Dialer.

Automatic Dialer is the applications developed which a list with data from several users destination is passed to the Gateway in order to be contacted. The Gateway will be responsible for making connections and in accordance with the response, connect the attendants to the Users "B".

This kind of application is widely used in corporations that perform credit recovery of its delinquent customers. This application can be used in cases of aggressive voice marketing merge for existing products, polls to vote, research on customer satisfaction, among others.

As regards the possibility of adding services to the specific company, this application is an example of the substantial amount of options that are outlined when adopting this kind of platform that integrates the world of telephony with the world of IT. With a single application can occur a variety of services to be provided by the corporation.

Below is the Fig. 3 that shows the flow of operation.

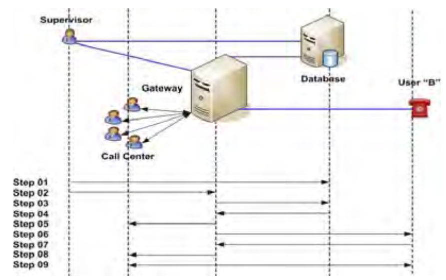


Fig. 3 Application – Automatic Dialer

In Step 01 of this flow, the Supervisor performs a load on the database, with list of names, telephone numbers and ID code of each User Destination. It is noteworthy that in the overall business where we implemented this solution, the average size of these lists dial is around 3000 contacts.

In Step 02 the supervisor start the application.

Step 03 is the time that the Gateway will ask the database a list of Target Users to be contacted.

In Step 04 the ratio is returned.

In Step 05 the Gateway will hold an investigation to see if there are attendants free in Call Center.

The sixth stage is the time when the gateway starts dialing according to the data contained in the database and already stored in memory.

In Step 07 the status of the User "B" is returned to the Gateway. The possibilities for this state are: answered, busy and unavailable. In this flow we are describing the status answered.

In Step 08, the Gateway forwards the call to the Callcenter and in accordance with the strategy used, an operator will answer the call.

When the call is established, occur the Step 09.

The next application that we will describe is Callback.

The main function of this application is to provide the return of the calls to the mobile phones that need to communicate and pay small financial rates.

The provision of such service becomes profitable for internal users within companies, specially here in Brazil that the mobile telephones rates still high.

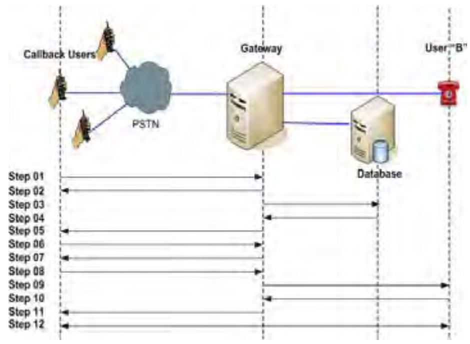


Fig. 4 Application – Callback

In Step 01 of this flow, the Callback User performs a call for a specific number of telephone connected to the Gateway.

During Step 02, the Gateway records the number of the Callback User and sends a signal to disconnect him.

Step 03 and Step 04 is when the number of the Callback User is checked and validates their existence inside the database.

In Step 05, the Gateway will performs the dial to the Callback User, previously validated by the confirmation of their existence in the database.

Step 06 is the confirmation of attendance by the Callback User.

In Step 07, the Gateway sends a dial tone to the Callback User.

The Callback User receives a dial tone, sends the number of the User "B" to the Gateway. This is Step 08.

In Step 09 the Gateway dial to the User "B" and awaits response from their state.

In Step 10, the User "B" responds with a state of 'not in use'.

In Step 11, the Gateway informs to the Callback User the state of User "B" and waiting for attendance by the User "B".

Step 12 is the point in time when the Callback User and User "B" remain connected.

Now we describe the tasks performed at the transport layer and topology of tests.

For the protocols(ISDN, R2Digital, SIP and IAX) we only needed to configure and integrate the respective libraries with Asterisk Core. On this layer, we used the open library chan_ss7 to validate the signaling SS7 Over IP to our Gateway. In chan_ss7 we needed to change the source code to perfect working of the signaling.

The complete library consists of the following modules: chan_ss7.c, l4isup.c, mtp.c and tranport.c.

In this library, the main change made was to correct the declaration of certain kind of variables that were defined as 'long int' and were changed to the 'int'. The other changes was the definition of data structure that not existed and we needed to create and set up.

We also observed errors regarding the path of certain files in the directory tree of the server.

This library is the only in which the installation was performed only in the laboratory, unlike the others that were performed in corporate environments.

The integration tests performed with the Gateway were divided into two distinct groups, namely: laboratory tests and enterprise environments test.

Other topics of changes made in this module are correlated with the integration between it and the native library of Asterisk core, namely DAHDI.

The main work of correction was performed by removing a portion of the code that occurred redundant definitions of functions and variables.

The feasibility of performing tests with SS7 Over IP in an environment of integration with the carriers was not possible due the unwillingness of the circuits for such purposes.

Below in Fig. 5 we show the topology used in laboratory to validate the signaling SS7 Over IP.

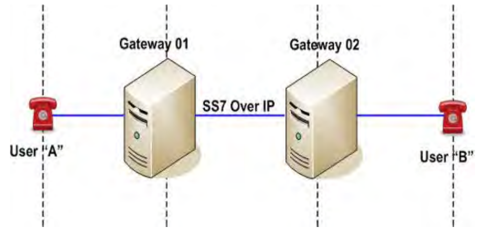


Fig. 5 SS7 Over IP Test Topology

In corporate environments tests, we validate real-world scenarios, the operation of the applications developed and the following protocols: SIP, IAX2, ISDN and R2Digital. The tests with traditional carriers were ISDN and R2Digital and with the current entrants in the emerging market for VoIP and IP Telephony, were the protocols SIP and IAX2.

Following Fig. 6 aims to summarize the scenarios for integration in corporate environments tests.

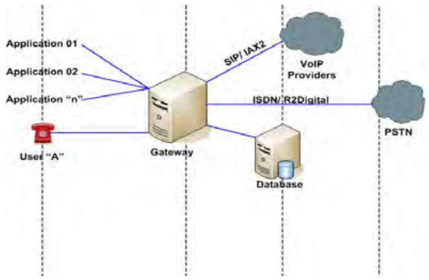


Fig. 6 Corporate Environment Test Topology

IV. CONCLUSIONS

We concluded that is possible to develop a gateway that provides integration between IP-PSTN through the open libraries and open platforms. This gateway permit access to a voice telecommunications architecture that has strong adherence and similarity to the functional NGN architecture.

The interconnection with telecommunications operators through the use of various protocols and the flexibility to develop new applications in a modular fashion allows us to confirm the actual adherence of the Voice Gateway to the stage of NGN.

To test the signaling SS7 Over IP with carriers and develop new applications for corporate environments is a good place to continues with this work.

REFERENCES

[01] K. Hechmi, G. Jean-Charles, "IMS for Enterprises", IEEE Communications Magazine, July 2007.

[02] Hyeong Ho Lee, "Signaling architecture and protocols for the Next Generation Network", Advanced Communication Technology, February 2009.

[03] S. Mittal, C. Dipanjan, S. Goyal and M. Sougata, "SewNet – A Framework for Creating Services Utilizing Telecom Functionality, April 2008.

[04] NetIQ, "How to Succeed on Your Journey to Unified Communications", September 2008.

[05] de Souza Pereira, J.H., Guilherme, J., Rosa, P.F, "Development of MGs in a next generation network with MEGACO/H.248 support", 12th IEEE International Conference on, November 2004.

[06] Tekelec, "SS7 Over IP – Standalone Signalling Gateways versus Integrated and Distributed Signalling Functions", 2007.

[07] Knightson, K.; Morita, N.; Towle, T. "NGN architecture: generic principles, functional architecture, and implementation", IEEE Communications Magazine, October 2005.

[08] Vaaramaki, T.; Puttonen, J.; Fekete, G. "Next Generation Network Related Standardization - Enablers for the Convergence", Fourth Advanced International Conference on, June 2008.

[09] ITU-T FGNGN-MR-0065.

[10] ITU-T Recommendation Y.2001, December 2004.

[11] Chae-Sub Lee; Morita, N."Next Generation Network Standards in ITU-T", BCN 2006 , The 1st International Workshop on, April 2006.

[12] Schreiner, F.; Wahle, S.; Blum, N.; Magedanz, T."Modular exposure of next generation network services to enterprises and testbed federations", ICCE 2008, Second International Conference on, June 2008.

[13] Ayatollahi, Z.; Sarukhani, S.; Fayazi, F.; Roknabady, Z.A.; Madani, A. "Interoperability problems in Next Generation Network protocols", K-INGN, First ITU-T Kaleidoscope Academic Conference, May 2008.

An Algorithm for Route Selection on Multi-sink Wireless Sensor Network Using Fuzzy Logic

Líliam Barroso Leal #1, Marcus Vinicius de Sousa Lemos #2, Raimir Holanda Filho #3

 # Mestrado em Informática Aplicada, University of Fortaleza
Av. Washington Soares, 1321, Edson Queiroz, CEP 60.811-905, Fortaleza CE, Brazil

1 liliam@edu.unifor.br
2 marvin@edu.unifor.br
3 raimir@unifor.br

Abstract—In this paper, we propose a new algorithm for route selection on Wireless Sensor Network (WSN) which allows communication between sensor nodes and multiple sinks. For that, the sensor nodes use a fuzzy inference machine in order to select the most appropriate sink. This inference will take into account some network characteristics such as: energy, number of hops, collisions, and delay. Simulations have shown that our algorithm is feasible not only in terms of routes selection, but also in terms of energy consumption.

I. Introduction

In the last years, there was a considerable increase in research involving wireless sensor networks (WSN), due to its applicability in many areas such as security, health, agriculture, among others. The WSN comprise a large number of small and intelligent sensor nodes capable of detecting and transmitting characteristics from the physical environment [1]. Intelligent sensors can be defined as a small circuit board containing an integrated sensor capable of processing and data communication [1].

Wireless sensor networks have unique characteristics very different from traditional networks, especially regarding the number of nodes present in the networks and constraints in some resources like energy, processing power, memory and bandwidth. These variables make impractical the re-use of most of the algorithms developed for traditional computer systems, since the design of any solution for sensor networks must deeply consider the consumption of energy and computational constraints.

The sensor nodes are deployed in a geographical area in order to observe events, such as atmospheric pressure, temperature, level of pollution, etc. These nodes can be randomly deployed or following some topology (which, in most of the cases, depend of the network designer and the area to be monitored). The sensors nodes, after sensing some event, must be able to send a report to a particular node, called sink (base station), which will analyze the received data and, according to the network goal, make some kind of decision.

Some of the most common communication paradigms in WSN involve communication of multiple sensors placed in an area of observation, reporting information to a single sink (many-to-one). In [2] it was shown that a single sink is a bottleneck in a network, especially for real-time applications. Thus, a multiple sink approach was proposed in order to overcome that problem. In this new paradigm, multiple sensors can communicate with multiple sinks (many-to-many).

The main goal of the algorithm proposed in this paper is to promote communication between sensor nodes and multiple sinks, where each node will select the best route to an appropriate sink node, localized somewhere in the network. This selection will be performed by a fuzzy inference machine. We have chosen this scenario because of the increasing popularity of applications which require networks with multiple sinks, for example, applications that use wireless multimedia sensor network [3], actuator nodes [4] and other applications in which there may be more than one base station. For example, a medical application [5], in which inside a hospital we could have sensors attached to patients and the doctors would be equipped with portable computing devices such as smartphones, which would act as sinks. Such scenarios require more appropriate solutions, which make the existing proposals, for networks that have a single sink, inefficient.

In our solution, the sensors will be able to select, wisely, the best route in a given time, based on some network characteristics, such as distance to the sinks, the energy from the neighbor nodes, among others.

The remainder of this paper is organized as follows: in Section II, we describe some related works; we present our proposed algorithm in Section III; in Section IV, we perform the simulation of our algorithm; in Section V, the results are depicted and finally, in Section VI we show the conclusions and future works.

II. Related Works

In this section, we will describe some related works which use a scenario of multiple sinks, as well as problems related to the approach of using a single sink.

In [2], we have a wireless sensor network with multiple sinks arranged in clusters, where each sink monitors a region with multiple events. The authors propose the use of a management system database, distributed in each sink of the network and require two changes on the routing protocol: the first is that the protocol should allow the creation and storage

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_103, © Springer Science+Business Media B.V. 2010

of multiple routes for the spreading of a message, and the second requires that the network provides QoS, because it works with the prioritization in the delivery of messages for notification of event. Once an event of interest is detected by a node, notification message is sent toward the sink. Paths are created and stored in each clusterhead (CH): each clusterhead stores information about the event and the ID of the cluster from which the message was received.

The research in [6] proposes a scheme based on fuzzy logic to select a better way and increase the survival time in MANETs. In that work, it has been mentioned that the survival of the network can be achieved when security measures, preventive, reactive and tolerant lines of defense, work together. This proposal presents the following contribution: a scheme of selection of adaptive path independent of protocol. Furthermore, the correlation of characteristics of network security and discretion are made through the use of fuzzy logic to improve the path selection minimizing security limitations found in routing protocols. The process of selecting path involves three phases: data collecting, the process of fuzzy inference and finally the selection of adaptive path.

The work in [7] investigates use of fuzzy inference in order to make the nodes select the best path in a flat wsn with only are sink. The fuzzy inference system selects the best candidate node basead on some metrics, like, for example energy. To carry out this goal some fuzzy rules are implemented, which use as metrics: the distance from the candidate node from the origin, the distance from the candidate node to the shortest path, energy and link usage. One of the disadvantages of this protocol is the fact that it does not address the use of multiple sinks, which makes it vulnerable to the problems mentioned above, such as bottleneck, and a much higher dissipation of energy of the nodes near the sink.

Our proposal differs from the related works, because it was developed for wireless sensor networks with multiple sinks and makes it possible to choose the best routes for sending data to a sink. We propose an algorithm that works with the routing protocol in order to assist a node in the process of selecting the best path among the possible routes in a given time. This process of route selecting is made by the use of a fuzzy inference system. Every route will be analized and classified regarding some criteria such as energy and number of hops. Whenever a node has to send some data, it will take the best classified route. Hence, the network life will increase as bad routes (route with little energy and/or much interference) will be avoided.

III. THE PROPOSED ALGORITHM

In this section, we will analyze in detail the proposed algorithm. Then, we will discuss general aspects of the algorithm in sub-section A, while in sub-section B, the selection phase of the routes with the aid of fuzzy inferences will be presented.

A. General Aspects

In our proposal, we consider that the sensors are placed uniformly in the region to be observed, and we also consider multiple sinks arranged in order to maintain coverage throughout the area. The sinks are devices with features far more advanced than the sensor nodes, having no limitations of power.

Each sink located in the network, will be responsible for receiving the notification messages of events at any point of it, for this, each sensor node will, through a fuzzy inference machine, select the most appropriate sink for it in that moment, that is, which the best route is taking into account the network characteristics, such as energy, number of hops, collisions and delay. We should emphasize that this study considers only some of these characteristics, but we can add new features according to the needs of the network.

It is important to emphasize that we need to select previously a routing protocol, because our solution works in conjunction with it. The choice of the routing protocol will be made according to the needs of the network designer, and it is only required that this protocol be capable of providing multi-path.

B. The proposed solution

We will now discuss the steps used in our algorithm to choose the best path for data transmission over the network. Table I shows the notations used, while Algorithm 1 presents the proposed algorithm.

TABLE I
NOTATIONS OF THE PROPOSED ALGORITHM

βn	The beacon message generated by the base station n
A :	The operation after the ":" The node is held in
$A \rightarrow * : \beta n$	A passes the message βn through broadcast
$A \rightarrow N : \beta n$	A passes the message βn to node N
$x \leftarrow y$	The value y is stored in the variable x

After the deployment, the nodes are initialized and the sinks generate a list containing the neighboring nodes (line 1 of Algorithm 1). This list is used by the sinks to send a beacon message βn. This process is depicted in lines 1 to 4 of Algorithm 1. The beacon message sent by the sinks is propagated to the network through the neighboring nodes. Upon receiving a beacon message, a sensor node A will check the following fields:

- Energy: this field is used to store the lowest energy level found in the path.
- Station_base: indicates the sink from which the beacon message originated.
- Number_hops: distance in hops from the origin sink of the message to this sensor node.
- Node_predecessor: identifies the node from which the node A received the message.
- Last_sequence: identifies the beacon message. Each time the sinks require a route rebuilding, a new beacon message is generated and the last_sequence field will be

incremented, so that the nodes can distinguish the current message from the messages sent previously.

Algorithm 1:

```
1 :  Base Station n: generates V list with neighboring nodes
2 :  Base Station:    n belonging to V
3 :     Base Station n: generates (βn)
4 :     Base Station → n : βn
5 :  A: receives ( βn )     ( βn ) transmitted by neighboring
          nodes
6 :     A: energy ← read field energy ( βn )
7 :     A: base_station: ← read field base_station ( βn )
8 :     A: number_hops ← read field num_hops ( βn )
9 :     A: node_predecessor ← read field node_predecessor
          ( βn )
10 :    A: last_sequence ← read field sequence_broadcast(βn )
11 :    A:  if A does not use the route predecessor
12 :          A: discard βn
13 :          A: exit
14 :    A: end-if
15 :    A: sequence_current = recover
          sequence_pkg_received  (node_predecessor)
16 :    A: if sequence_current >=  last_sequence
17 :          A: discard βn
18 :          A: exit
19 :    A: end_if
20 :    A: FL ← calculate FL (energy, number_hops)
21 :    A: update_FL_route (node_predecessor, FL)
22 :    A: if energy(A) < energy( βn ) then
23 :          A: store energy (A) in the field energy of βn
24 :    A: end_if
25 :    A → * : βn
26 :    A: end_if
```

If a node D receives a beacon message whose value of field *node_predecessor* is different from some of the neighboring nodes it uses as a current route, it will drop this message, as illustrated in Figure 1. This measure is adopted so that node D does not have to calculate a fuzzy level for a node it does not use as a route.

Aiming to prevent old messages from keeping on traveling through the network, some routines are set according to lines 15 to 19 of the algorithm presented. Whenever node A receives a beacon message, it checks the field *last_sequence* to verify whether it refers to a new beacon message, or if it is just the spread of a message received earlier. The next step is the calculation of fuzzy level (FL) of the route; this parameter allows the node to select the best route in a given moment, according to the last beacon message received. Each route that a node can use, it is associated with a FL. Routes with a higher value of FL are considered preferential. The calculation is carried out using the FL and fuzzy logic, and in our algorithm we have considered as parameters the linguistic values energy and number of hops. In the sub-section below, we will explain this calculation in more detail. Finally, the node A tests if its energy is lower than the energy stored in the received message, if so, it updates the field *energy* of the beacon message with the value of its remaining energy, otherwise it goes to the next step and relays the message through the network.

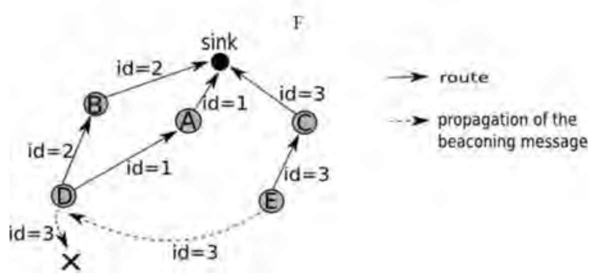


Fig. 1. Disposal of Beacon Message

In our work, we have considered that each sink node sends beacon messages in order to make the sensor nodes update their fuzzy levels in network, within a window of time α, where α is a value to optimize the lifetime of the network. The sensor nodes in the network check their own energy for each time window θ. We have also set a safety limit of the amount of remaining energy in the node, represented by . This limit makes the node able to report a poor energy condition at any time, without the need to wait for the next beacon message from the sinks. The value of is informed by an alert message propagated through the nodes with hop = 5, avoiding the wasting of energy by unnecessary broadcasts. All values assigned to θ, and hop may also be defined by the network designer.

C.The Fuzzy Inference Machine

Fuzzy inference can be defined as a process of mapping a given input into an output through the use of fuzzy logic. This mapping is carried out by rules that define the output associated with a particular input [6]. For example, if the energy condition of a node A is bad, then do not forward packets to this sensor node. These rules have as input values defined by the linguistic variables, which in this work are represented by energy, figure 2 and number of hops, figure 3 and they can be altered according to the need for the network designer. Each linguistic variables has a set of linguistic values (fuzzy values) associated to it. As an example of linguistic values of energy we have low, medium and strong. Each linguistic variable has its membership function which will convert crisp values (real values measured from the environment) into fuzzy values. And both (rules and membership functions) are responsible for the behavior of our algorithm regarding the routing decision. It is important to point out that in fuzzy sets, the values can range from 0 to 1, with 1 indicating the value of absolute truth and 0 absolute false.

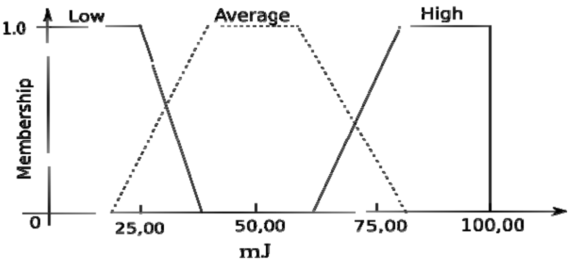


Fig. 2. Function Energy

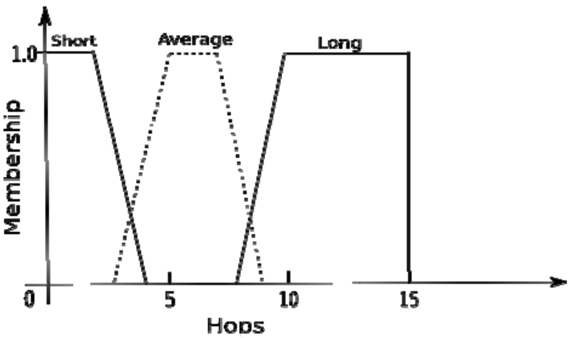


Fig. 3. Function Path Length

IV. SIMULATION

This section presents comments on the simulations performed to validate the proposed routing algorithm.

A. *Simulator*

In our experiments, we used the simulator Sinalgo [8]. Sinalgo is a framework, written in Java, for testing and validation of networks algorithms. Unlike others tools such as Network Simulator 2 [9] which allows the simulation of various network stack layers, Sinalgo focuses on the verification of algorithms and abstracts from lower layers (like physical, link, network, among others).

B. *Network Features*

We will present the features of the network as follows:

1) The network topology: the simulated network is flat, fixed and contains only two types of nodes: sensor nodes and sink nodes. Each sensor node has a single identification and a fixed radio range.

2) Used scenarios: in our simulations, we have used four scenarios with a similar environment. All scenarios are comprised by 100 sensor nodes uniformly distributed along the network, covering an area of 700 x 700 m2. We simulated the following scenarios:

- Scenario 1: without multiple sinks and without fuzzy logic. In this scenario we have one sensor node transmitting data to one sink.
- Scenario 2: without multiple sinks, without fuzzy logic, and with route balancing. In this scenario, whenever a sensor has to send packets toward the sink, it will balance the packets between their routes. For example, a node *X* has three routes represented by the numbers 1, 2, and 3 and has 4 packages to send to the sink, represented by the letters A, B, C, and D. The node will forward them as follows: The package A will be forwarded using route 1, the packet B using route 2, the package C using route 3 and the package D will be sent through route 1.
- Scenario 3: without multiple sinks and with fuzzy logic. In the scenario, one sensor node transmits data to one sink, using together with the routing algorithm, the proposed solution with the use of fuzzy levels to choose alternative routes.
- Scenario 4: with multiple sinks and fuzzy logic. In this last scenario we have one sensor node transmitting data for two sinks using fuzzy levels for choosing the best routes.

3) Routing Algorithm: For the communication between nodes and sinks, we have created a variation of the Protocol Destination-Sequenced Distance Vector (DSDV), [10] in which all the sensor nodes will transmit the data collected to a sink. The system starts without any knowledge about the identity or the topology of sensors which are present. Each node knows only its own identity. The sinks broadcast routing update messages periodically. Devices that are in direct range of the sink, upon receiving this message, update their route tables. These nodes perform then a broadcast of a new routing update message to any device within its range (Figure 4a), indicating that there is a path to a sink through them (nodes which have sent the broadcast). To implement the multi-path routing scheme, the devices will store the first three routing update messages which they receive (Figure 4b).

The algorithm used, although not the most efficient, has been shown satisfactory since, in this work, we have only analyzed the behavior of our application under a multi-path routing protocol, leaving for future work the analysis of our solution using other routing protocols.

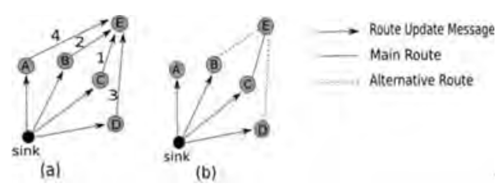


Fig. 4. Routing Algorithms Used

V. RESULTS

A. *Scenarios according to time*

The results of the proposed algorithm are presented regarding three metrics according to the experiments. Below, we have defined our metrics:

- Time of death of the first node: In Figure 5, it is expressed the time of death of the first node in the network. Here, we want to analyze how long our proposed work makes the nodes alive.
- Time of death of the network: Here, we investigate how long the network stay alive, that is, when the network could no longer maintain the necessary communications, Figure 6.
- The numbers of the sink nodes: Here, in this metric we wants to see how the numbers of sink nodes can influence our work, Figure 7.

Fig. 5. Death Time of the First Node

Fig. 6. Death Time the Network

We can realize from Figure 5 that the scenario 1 (without fuzzy logic and without multiple sinks) has been shown more fragile as regards the life of a node which participates actively in a route in use. We can also infer that in the scenario 2 (without multiple sinks, without fuzzy logic, and with route balancing), represented by the dark gray column, showed a gain of approximately 42% compared to the scenario 1, black column, because the route balancing. Scenario 3 (without multiple sinks and with fuzzy logic), illustrated by the light gray column, compared with scenario 1 showed an increase of 106%, because of the possibility of using alternative paths inferred by FL. Finally, on scenario 4 (depicted by a white column), we have the representation of the proposed solution with the use of fuzzy logic and multiple sinks, in which the gain, compared to the scenario 1, is approximately 234%. This gain was expected because of multiple sinks, increasing the number of alternative routes by which the node can use for data transmission and the fuzzy inference which make the nodes avoid the bad routes.

In Figure 6, we have the representation of the time of death of the network, time which it prevents the transmission of data due to the death of some nodes in the path of routes. The black column represents the scenario 1 without fuzzy logic and multiple sinks, in which the time of death of the network occurred after the death of the first node of the network, because this scenario does not have multiple routes and multiple sinks. In the scenario 2, shown by the dark gray column, the network was active for a longer time, approximately 222% more than the scenario 1; this gain was achieved by the use multiple routes. The light gray column depicts scenario 3, where the network was active for a longer time, approximately 330% more than the scenario 1; this gain was achieved by the use of fuzzy logic, although the feature of multiple sinks has not been added. We have used this scenario to assess the choice of routes from the node, and thus determine whether the fuzzy rules created have reached expectancy. Finally, the white column shows the use of fuzzy logic and multiple sinks, in which we have obtained a time of activity of the network higher than the previous scenarios.

We also evaluated the behavior of our solution according to the amount of sink nodes that we have on the network, while still maintaining a satisfactory result. Therefore, gradually add more sink nodes in the environment and simulate their behavior as can be seen in Figure 7. We realize that to some extent the number of sink nodes influence positively the time of death of the first node, however, as more sink nodes are added, the greater the number of beacon messages traveling on the network, affecting the life nodes.

Fig. 7. Influence of the number of sink nodes

B. *Window of time α*

As mentioned previously, in our algorithm we consider that each sink expect a time window α for sends beacon messages in order to make the sensor nodes updates its fuzzy levels. It was performed simulations with different values of α and we evaluated the performance based on death time of the first node of the network (Figure 8). We can see that the values near to the optimum for this set are between 500 and 700 seconds.

Fig. 8. Values for Time Window α

VI. CONCLUSION

This work proposed an algorithm for path selection using fuzzy logic on WSN with multiple sinks. The main goal of the algorithm is to extend the network time life and this can be achieved by making the nodes avoid bad routes while sending packets through the network. We used fuzzy logic to build up a knowledge base that makes possible the process of finding the best routes

Results obtained from simulations have shown that our algorithm is feasible, increasing up to 479% the lifetime of the network, since several routes can be used for message forwarding.

In future works, we intend to analyze the behavior of algorithms in more complex scenarios, and select new linguistic variables for fuzzy inference.

VII. REFERENCES

[1]Karl, Holger and Willig, Andreas, *Protocols and Architectures for Wireless Sensor Networks*, John Wiley & Sons, Ltd, 2005.

[2]Boukerche, Azzedine and Martirosyan, Anahit, *An Energy Efficient and Low Latency Multiple Events' Propagation Protocol For Wireless Sensor Networks With Multiple Sinks*, In: PE-WASUN'07, 2007, Chania, Crete Island, Greece.

[3]Akyildiz I. F., Melodia T., Chowdhury K. R., "A Survey on Wireless Multimedia Sensor Networks" Computer Networks (Elsevier), Vol. 51, no. 4, pp.921-960, March 2007.

[4]Pietro Ciciriello and Luca Mottola and Gian Pietro Picco, *Efficient Routing from Multiple Sources to Multiple Sinks in Wireless Sensor Networks*, In

Proceedings of the 4th European Conference on Wireless Sensor Networks (EWSN), 2007.

[5]Kim, Yung Bok and Kim, Marie and Lee, Yong Joon, *COSMOS: a middleware platform for sensor networks and a u-healthcare service*, SAC '08: Proceedings of the 2008 ACM symposium on Applied computing, 2008.

[6]Michele Nogueira Lima and Helber W. da Silva and Aldri L. dos Santos and Guy Pujolle, *Survival multipath routing for MANETs*, NOMS, 2008.

[7]Mohammad Abdul Azim and Abbas Jamalipour, *Optimized Forwarding for Wireless Sensor Networks by Fuzzy Inference System,* Auswireless Conference, 2006.

[8]Sinalgo (2007). http://dcg.ethz.ch/projects/sinalgo/

[9]NS2 (2008). http://nsnam.isi.edu/nsnam/index.php/main_page.

[10]Perkins, C. E. and Bhagwat, P. (1994). Highly dynamic destination-sequenced distancevector routing (dsdv) for mobile computers. SIGCOMM Comput. Commun. Rev., 24(4):234-244.

Mining Mobile Sequential Patterns in Wireless Cellular Networks

Joshua G. Bradley and Sherif S. Rashad
Department of Mathematics and Computer Science
Morehead State University
Morehead, KY 40351
jgbrad01@moreheadstate.edu
s.rashad@moreheadstate.edu

Abstract— **In this paper, we have taken concepts of data partitioning methods and modified the well known algorithm SPADE (Sequential PAttern Discovery using Equivalence classes), which we have implemented over a mining model, called mining mobile sequential patterns, and called MobileSPADE. From this, we have devised two innovative forms, called Fixed-MobileSPADE algorithm and Dynamic-MobileSPADE algorithm, which will mine for mobile sequential patterns from either fixed-length or dynamic-length transactions. In mining for mobile sequential patterns in a mobile commerce environment, we used tower ID data extracted from a dataset that was constructed and managed by the Massachusetts Institute of Technology (MIT) Media Lab: Reality Mining Project, which logged objective data from 100 mobile device users for a period of nine months, to represent mobile user's movements. Prediction analyses results were used to measure the performance of Fixed-MobileSPADE and Dynamic-MobileSPADE algorithms on the mining mobile sequential patterns model. Results show that as the quantity of items that are taken into consideration for mining mobile sequential patterns increases, the prediction accuracy correspondingly increases.**

I. INTRODUCTION

In a world where mobile devices are becoming more popular and one of the leading factors to an end-user in choosing a mobile network carrier is cell coverage for a particular area, advancements in wireless communication technologies and routing protocols is often advantageous to a network carrier in order to gain more customers. When studying a cellular network, it is often best to view the network as a grid of cells where each cell represents the coverage area for a cell tower, similar to the grid in Figure 1. Due to the extent of applications now available for mobile devices, there have been many studies on Location-Based Service (LBS) in the past couple of years, but one important research issue that is becoming prevalent in upcoming network technologies is the tracking and prediction of a mobile user's movement pattern in a cellular network. Currently, a mobile network carrier uses the following strategy to maintain cell coverage and introduce a node to its network. Using Figure 1 as an example, when a mobile user is in cell A, the mobile network carrier will distribute and allocate resources required by the mobile user evenly throughout the surrounding cells. This allows for the network to be readily prepared to continue the node's data connection if the mobile user continues using the mobile device and transfers to an adjacent cell. Disadvantages to this strategy include producing unnecessary overhead by distributing resources to cell towers that will not be used by the mobile user, thus not allowing the mobile user full employment of all available resources. One alternative solution that has been extensively researched has been the mining for frequent sequential patterns. In this paper, frequent sequential patterns refer to the frequent *movement* patterns of a mobile user. Finding these frequent sequential patterns could allow for a network carrier to calculate with some probability where a mobile user will go next, based upon his/her past movement history. Using this innovative mobile sequential pattern model and looking at Figure 1 again, when a mobile user resides in cell A, the network carrier could allocate more resources of a certain service requested by the user to the cell most likely predicted to contain the end user. For example, based upon a mobile user's past history, if the mobile user is calculated to have a 75% chance of going from cell A to cell D, then the network can allocate/distribute more of its resources to cell D as compared to the surrounding cells, in order to provide better service to the end-user via a more efficient use of its resource allocation methods and routing protocols.

Figure 1 A sample grid of cells in cellular networks

The remainder of this paper is organized as follows. The next section reviews the preliminary concepts needed to comprehend the rest of this paper and section 3 will discuss the concept of data partitioning methods and the original SPADE algorithm. The dataset provided by the Massachusetts Institute of Technology Reality Mining Project is discussed in section 4. Section 5 discusses algorithm MobileSPADE for discovering frequent sequential

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_104,

patterns in a cellular network and its two derivatives, algorithm Fixed-MobileSPADE and algorithm Dynamic-MobileSPADE. Section 7 presents the experiment results obtained from algorithm Fixed-MobileSPADE and algorithm Dynamic-MobileSPADE and section 8 discusses the conclusion on the experiment results.

II. Related Work and Preliminary Concepts

In this section, we analyze an experiment conducted in [1] on the mining of frequent itemsets via a data partitioning approach which employs a modification of the algorithm in [2]. We also discuss the SPADE algorithm and which aspects were modified/ enhanced during the development of Fixed-MobileSPADE and Dynamic-MobileSPADE.

In [1], Nguyen and Orlowska enhance the data pre-processing approach of the Partition algorithm [2] to increase the performance of frequent itemset computations. A series of data pre-processing methods such as the introduction of new clustering and similarity algorithms were applied across the data to show that the performance of the algorithm discovering frequent itemsets increases by means of data partitioning. The Partition algorithm in [2] follows the concept that a fragment $P \subseteq D$ of the database is a subset of transactions within the database. With reference to an itemset, *local support* is defined as the fraction of transactions containing that itemset and *local candidate itemset* describes an itemset meeting the minimum support for the whole database. *Global support* and *global frequent itemset* are calculated similarly except with respect to the whole database. The principle behind [1] is that each fragment is calculated in the main memory one at a time.

In [3], Zaki presents SPADE, an algorithm designed for the accelerated discovery of sequential patterns. Previous algorithms have required multiple scans over a database for their calculations, but SPADE utilizes combinatorics to decompose the original problem into smaller sub-problems, which are individually solved in the main memory and later combined.

III. Mobility Dataset and Preliminary Concepts

It is important to discuss the mobility dataset used to obtain performance tests results on Fixed-MobileSPADE and Dynamic-MobileSPADE algoritms. The dataset used in this research is a result of the reality mining project carried out by the MIT Media Lab [4]. This dataset was collected over a nine month period and involved the use of 100 cellular devices [4, 5]. As explained in [4] mobile device logging software was designed and installed on 100 mobile devices, which recorded critical information about the mobile user, such as proximity, location, services used, length of calls made and received, type of data used, etc onto the device's memory card. The data for all 100 mobile users was collected and organized into one dataset. This dataset has been of great interest recently to researchers involved in mobile data mining and the discovery of frequent sequential patterns, because this is one of the first research projects that successfully collected an extensive amount of objective data, as opposed to subjective data. In this paper, objective data refers to the data collected by the mobile device logging software. Subjective data refers to previous methods used to test algorithms, which include creating a virtual mobile commerce environment and issuing survey's to record relative location.

There is great potential for this dataset and other similar datasets to revolutionize the whole concept of social networking and advertising. Use of this dataset provides researchers with a data model that represents objective human behavior. Prospective exploits of this dataset include the testing and modification of current cellular network algorithms to improve their performance, re-defining current advertising techniques to make advertising targeted and more personalized to an individual mobile user, and developing innovative methods to find frequent sequential patterns.

Notation used throughout the paper is as follows:

- Let $T = \{t_1, t_2 \ldots t_n\}$ be a set of n transactions over data set D.
- Transactions can be of fixed or variable length, depending on the algorithm used.
- Each transaction consists of a sequentially ordered set of items $\{i_1, i_2, i_3 \ldots i_m\} \subseteq T_n$, referenced as tower ID numbers in this paper.

The goal of Fixed-MobileSPADE and Dynamic-MobileSPADE algorithms is to generate all the discovered frequent itemsets that meet a user-defined support threshold and confidence threshold.

IV. Algorithms for Mining Frequent Mobile Sequential Patterns

Similar to SPADE algorithm [3], algorithm MobileSPADE is an Apriori-based sequential pattern algorithm that utilizes characteristics of combinatorics and adapts a vertical data format of the original dataset. Furthermore, two defining characteristics of algorithm MobileSPADE lie within its ability to incorporate data attributes into a consecutive sequentially ordered list, based upon timestamp values, and its capacity to only consider itemsets of length two or greater *len(itemset)* ≥ 2, in order to find frequent sequential itemsets. When using the mobile dataset from the MIT Reality Mining Project, the timestamp is defined by the values in the "starttime" column of the dataset. A mobile user's movement path is defined as the tower ID utilized for each logged event. Table 1shows a sample of the collated dataset with the corresponding *starttime* and *celltower_oid* data columns for mobile user number 94 after sorting with reference to *timestamp* value. We can see the mobile user utilized *cell_tower(38)* at *timestamp(2004-07-23 12:22:34)* and then *cell_tower(39)* 22 seconds later at *timestamp(2004-07-23 12:23:56)*.

Table 1 A sample dataset with starttime and celltower_oid data columns for user number 94 after sorting with reference to starttime value

User 94	
Starttime	celltower_oid
2004-07-23 12:22:34	38
2004-07-23 12:23:56	39
2004-07-23 12:25:10	38
2004-07-23 12:26:15	39
2004-07-23 12:27:24	40
2004-07-23 12:28:36	38
2004-07-23 12:28:51	40
2004-07-23 12:29:14	39
2004-07-23 12:33:40	41
2004-07-23 12:34:37	40
2004-07-23 12:37:16	39
2004-07-23 12:38:22	38
2004-07-23 12:38:49	41
2004-07-23 12:40:21	39
2004-07-23 12:41:01	40
2004-07-23 12:48:37	41
2004-07-23 12:49:29	40
2004-07-23 12:51:50	38
2004-07-23 12:53:20	41
2004-07-23 12:55:18	40

The original SPADE algorithm was proposed by [3] and did not account for consecutive sequential order as a parameter when finding frequent itemsets. SPADE algorithm was designed to utilize all candidate subsets of a transaction when calculating the support and confidence of each itemset. In our problem, it is important to keep consecutive sequential order of the items (the networks cells in this case). It is important to note also that when calculating the confidence of each itemset , the item at the end of each transaction is not considered by algorithm Fixed-MobileSPADE, because there is no succeeding movement beyond that item (or that cell).

Using the sample dataset shown in Table 1 as an example, the Fixed-MobileSPADE algorithm and the Dynamic-MobileSPADE algorithm are described as follows:

A. Algorithm Fixed-MobileSPADE

Input: *Transaction database: D*; *transaction_length*; *support_threshold*; *confidence_threshold*
Output: frequent sequential itemsets

Begin

1) Sort database *D* in sequential order with reference to *timestamp* value (*starttime*). The result will be as shown in Table 1.

2) Divide database *D* into equal-length transactions; each transaction containing the user-defined *transaction_length* number of items. For example, if *transaction_length* = 5, then T = {t_1, t_2, t_3, t_4} where

   t_1 = {38, 39, 38, 39, 40}
   t_2 = {38, 40, 39, 41, 40}
   t_3 = {39, 38, 41, 39, 40}
   t_4 = {41, 40, 38, 41, 40}

3) For each transaction t_n, find all possible **sequential** subsets with length ≥ 2. Using the current example, a list of candidate sequential itemsets would include:
   t_1 = (38, 39), (38, 39, 38), (38, 39, 38, 39), (38, 39, 38, 39, 40)
   t_2 = (38, 40), (38, 40, 39), (38, 40, 39, 41), (38, 40, 39, 41, 40)
   t_3 = (39, 38), (39, 38, 41), (39, 38, 41, 39), (39, 38, 41, 39, 40)
   t_4 = (41, 40), (41, 40, 38), (41, 40, 38, 41), (41, 40, 38, 41, 40)

4) Calculate the support *support*(i_m) of each sequential candidate itemset when *support*(i_m) = *frequency_count*(i_m)/ *number_of_transactions* where *frequency_count*(i_m) = frequency of itemset i_m in database D and *number_of_transactions* = |D| / *transaction_length*. If *support*(i_m) < *support_threshold*, then remove the itemset from candidate sequential itemsets.

5) Calculate the confidence *confidence*(i_m) of each sequential candidate itemset when *confidence*(i_m) = *frequency_count*($A \cup B$) / *frequency_count*(A) where A = *itemset[0]* the first element of the itemset and ($A \cup B$) = *itemset[0:]* the set of all items in the itemset. If *confidence*(i_m) < *confidence_threshold*, then remove from candidate sequential itemsets.

End

B. Algorithm Dynamic-MobileSPADE

Input: *Transaction database: D*; *support_threshold*; *confidence_threshold*
Output: frequent sequential itemsets
Begin

1) Sort *Sort(D)* database by sequential order, with reference to timestamp value. *Refer to Figure 2

2) Sort T into t_n transactions by creating non-repeating sequences of itemsets. For example,

3) when using the database from Figure 3, T = {t_1, t_2, t_3, t_4, t_5, t_6, t_7} where:

   t_1 = {38, 39}
   t_2 = {38, 39, 40}
   t_3 = {38, 40, 39, 41}
   t_4 = {40, 39, 38, 41}
   t_5 = {39, 40, 41}
   t_6 = {40, 38, 41}
   t_7 = {40}

4) Consider each transaction as a sequential itemset, where the length of each itemset must be greater than two *len(itemset)* $\geq$ *2*. Using the current example, a list of candidate sequential itemsets will be as follows:

   t_1 = (38, 39)
   t_2 = (38, 39, 40)
   t_3 = (38, 40, 39, 41)
   t_4 = (40, 39, 38, 41)
   t_5 = (39, 40, 41)
   t_6 = (40, 38, 41)

5) Calculate the support *support*(i_m) of each sequential candidate itemset when *support*(i_m) = *frequency_count*(i_m)/ *number_of_transactions* where *frequency_count*(i_m) = frequency of itemset t_n in database D and *number_of_transactions* = |D| / *transaction_length*. If *support*(i_m) < *support_threshold*, then remove the itemset from candidate sequential itemsets.

6) Calculate the confidence *confidence*(i_m) of each sequential candidate itemset when *confidence*(i_m) = *frequency_count*($A \cup B$) / *frequency_count*(A) where A = *itemset[0]* the first element of the itemset and ($A \cup B$) = *itemset[0:]* the set of all items in the itemset. If *confidence*(i_m) < *confidence_threshold*, then remove the itemset from candidate sequential itemsets.

End

Algorithm Dynamic-MobileSPADE is similar to algorithm Fixed-MobileSPADE in that when calculating the confidence of each itemset, the last item in each transaction is not considered. One significant difference between algorithm Fixed-MobileSPADE and algorithm Dynamic-MobileSPADE is the greater number of sequential itemsets considered in algorithm Fixed-MobileSPADE. This occurs due to differences found in the candidate sequential itemset generation phase of each algorithm. Algorithm Fixed-MobileSPADE takes all sequential subsets of each transaction into consideration while algorithm Dynamic-MobileSPADE only considers each transaction as an itemset. Once both algorithms have generated all itemsets of length two or greater that meet the minimum support and confidence threshold, sequence rules can then be defined, based upon the sequential itemset results.

V. Experiment Results and Prediction Analysis

We conducted multiple experiments with algorithm Fixed-MobileSPADE and algorithm Dynamic-MobileSPADE, using different parameters, on the MIT dataset. These experiments are performed on a 2.5 GHz Intel CPU and 4096 MB of memory. There were many results collected, but because of the limited space and for the purpose of explanation, we will only use length-2 itemset results in this paper.

We tested Fixed-MobileSPADE algorithm over 1000 records from the MIT dataset and let *transaction_length* = 10, *confidence_threshold* = 20%, and *support_threshold* = 20%. We found that 11 length-2 itemsets met these parameters and are now considered to be frequent length-2 sequential patterns over 1000 examples by Fixed-MobileSPADE algorithm. We used also the same previous parameters over 2000 records from the dataset. The results show that six itemsets met these specifications and are now considered to be frequent length-2 sequential patterns over 2000 records by Fixed-MobileSPADE algorithm. We tested also the Dynamic-MobileSPADE algorithm using different values for *confidence_threshold* and *support_threshold* .

After finding the frequent sequential itemsets that meet the user-defined minimum support threshold and confidence threshold, experiments were conducted on the dataset to test for prediction accuracy of Fixed-MobileSPADE and Dynamic-MobileSPADE algorithms. Using a minimum support threshold and confidence threshold of 20% for user 94 in the dataset, results for Fixed-MobileSPADE algorithm are graphed in Figure 2, where using 200 examples yielded the best results. Using a minimum support threshold and confidence threshold of 5% for user 94 in the dataset, results for Dynamic-MobileSPADE algorithm are shown in Figure 3, where using 200 examples yielded the best results.

We can see that, overall, Fixed-MobileSPADE algorithm performed better for user 94 due to the fact that user 94 exhibits more frequent behavior and the Fixed-MobileSPADE algorithm considers more candidate sequential itemsets than Dynamic-MobileSPADE algorithm. We can say that Fixed-MobileSPADE algorithm was designed to perform best for a user who exhibits more frequent behavior, whereas Dynamic-MobileSPADE algorithm was designed to perform better under conditions where the user exhibits more sporadic behavior. This can especially be seen from the experiment results for the

Dynamic-MobileSPADE algorithm in Figures 3 and Figure 4, where the minimum support threshold and minimum confidence threshold have been lowered and the number of records considered by the algorithm increased. Even with these two parameters changed, prediction accuracy still did not outperform Fixed-MobileSPADE algorithm. Again, this occurs because algorithm Fixed-MobileSPADE takes more itemsets into consideration than Dynamic-MobileSPADE algorithm when determining candidate sequential itemsets.

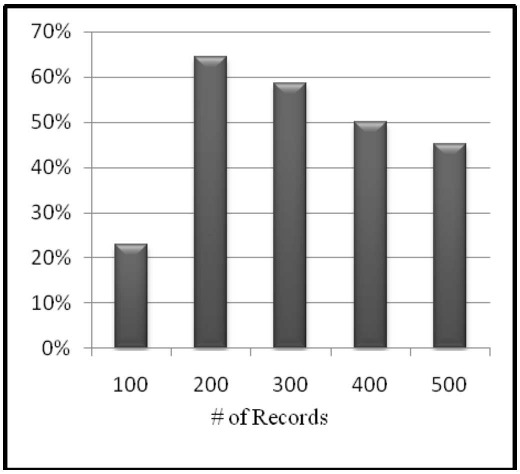


Figure 2 Prediction accuracy using Fixed-MobileSPADE for Support Threshold = 20% and Confidence Threshold = 20%

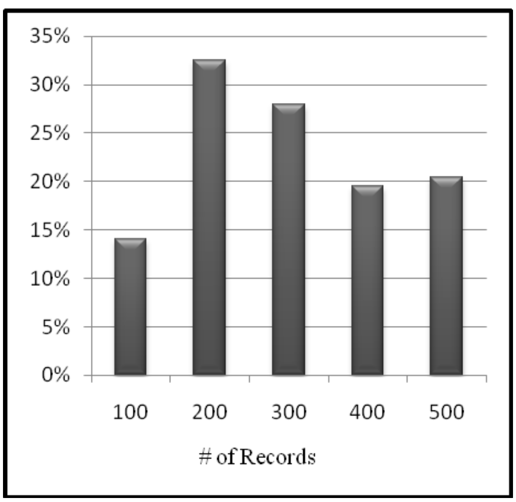


Figure 3 Prediction accuracy using Dynamic-MobileSPADE usforing Support Threshold = 5% and Confidence Threshold = 5%

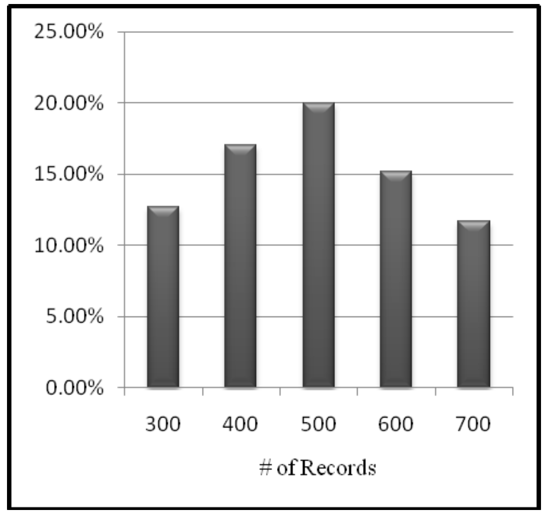


Figure 4 Prediction accuracy using Dynamic-MobileSPADE for Support Threshold = 10% and Confidence Threshold = 10%

VI. CONCLUSION AND FUTURE WORK

We have proposed in this paper two mobile data mining algorithms that can be used to support the wireless mobile networks. These algorithms are Fixed-MobileSPADE algorithm and Dynamic-MobileSPADE algorithm. These algorithms are modified versions of the SPADE (Sequential PAttern Discovery using Equivalence classes) algorithm. These algorithms were used to predict future movements of mobile users using mobility sequential patterns that are generated by Fixed-MobileSPADE and Dynamic-MobileSPADE algorithms. These algorithms were evaluated using the mobility data of 100 mobile users for a period of 9 months. These data were collected by the MIT Reality Mining project. Prediction analysis results show Fixed-MobileSPADE algorithm can be used to extract mobility sequential patterns of mobile users with frequent behavior while Dynamic-MobileSPADE algorithm can be used to extract mobility sequential patterns of mobile users with sporadic behavior. Future development of Fixed-MobileSPADE algorithm and Dynamic-MobileSPADE algorithm includes extensions to consider other factors when predicting a mobile user's future movements, such as time of day, services utilized in each cell, and an enhancement in the time series analysis currently undertaken by algorithm MobileSPADE.

REFERENCES

[1] S. N. Nguyen and M. E. Orlowska, "A Further Study in the Data Partitioning Approach for Frequent Itemsets Mining," in 17th Australasian Database Conf., pp. 31-36, Hobart, Australia, 2006.
[2] A. Savasere, E. Omiecinski, and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases," in Proc. Of the 21st Very Large Databases Conf., pp. 432 – 444, Zurich, Switzerland, 1995.

[3] Zaki, M., "SPADE: An Efficient Algorithm for Mining Frequent Sequences," *Machine Learning Journal*, 42(1/2), pp.31-60, 2001.

[4] N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems," Personal and Ubiquitous Computing, Vol 10, #4, pp. 255-268, 2006.

[5] B. Congleton and S. Nainwal, "Mining the Mine - Exploratory Social Network Analysis of the Reality Mining Dataset," School of Information, University of Michigan, Ann Arbor, Michigan.

[6] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. San Francisco, MK, 2006, ch. 8, "SPADE: An Apriori-Based Vertical Data format Sequential Pattern Mining Algorithm" , pp. 502-508.

Agile Software Development Methods and its Advantages

Dr.Soly Mathew Biju
University of Wollongong in Dubai

Abstract **–There is a high trend observed in the industry today towards adopting Agile software development method. This paper explains the principles of Agile software development approach and its benefits. It also exhibits how Agile software development method reduces the impact of frequently changing requirements throughout the development process. The paper will also describes the two agile development methods used in the IT industry-Extreme programming(XP) and SCRUM and the major difference between their work styles and processes. It highlights the advantages of Agile software development over traditional software development lifecycle.**

I. Introduction

Traditional software development methods (water-fall, spiral, Iterative) were very cumbersome and involved a lot of documentation. These methods were very rigid, well planned and hence provided very little or no flexibility. McCauley [4] states that "the underlying philosophy of process-oriented methods is that the requirements of the software project one are documented, they are locked and frozen before the design and software development process takes place". This process is not flexible enough to accommodate changes in the specifications at a later stage of the development. According to Highsmith and Cockburn [3], "what is new about agile methods is not the practices they use but their recognition of people as the primary drivers of project success, coupled with an intense focus on effectiveness and manoeuvrability. This yields a new combination of values and principles that define an agile world view".

Project development would prove to be a disaster if the customer is not sure of the final product requirements and can give the requirement only on an incremental basis. It follows the CCM (Capability Maturity Model), which is process based. The "Agile Method" is people based rather than plan based. In Agile software development the focus is on customer satisfaction. It is an incremental/iterative software development process which focuses on quick development of a working model.

The Agile software development method is based on the following principles:

- Individuals and interactions over processes and tools.
- Working software over comprehensive documentation.
- Customer collaboration over contract negotiation.
- Responding to change over following a plan.

The Agile Software Development manifesto was published by a group of software practitioners and consultants in 2001 [3].

II. EXTREME PROGRAMMING (XP)

A. Introduction and benefits.

"Extreme programming" is one of the commonly used Agile software methods [7]. It is used in the problem domains where the initial requirements are likely to change to a large extent.

It is based on the following assumptions:

- Be willing to do rework frequently and fast, as well as welcome customer feedback. It becomes a customer focused and customer driven project rather than a plan driven project.

- Initial software is developed based on the stories and requirements given by the user. This software will contain the most important, vital feature required by the user. This method is found to be very effective as it involves the customer from the very beginning. The customer writes the user stories comprising simple lines of text describing the features they want included in the first release of the working program. The acceptance test for the software is designed and based on these user stories. This will keep a check on whether all the features requested are implemented in the program.

- The developer then estimates the time for the development activity for which they will have to meet the customer, reconfirm and

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_105, © Springer Science+Business Media B.V. 2010

make detail note of the requirements. The first release should not take longer than three weeks.

- Hence this method consists of small release cycles for which the feedback of the customer/user stories is written which are then used for the next release of the software.

- The length of the increments is decided and is usually equal for all releases.

- The developers are advised to keep their code as simple as possible and the technically latest version which in turn minimizes the need for documentation.

- Requirements can be added anytime during the project development.

- New features can be added to the software as the next increments are provided.

- Before the next release the customer will have to prioritise the requirements for that release. The next release will focus on these requirements. All the contents of the release are subject to change except for the new feature implemented [9].

- If analysis of the requirements shows that the increment will take longer than the decided increment length, it is broken down into smaller increments.

- Each increment will have the developer working on the analysis, design code integration and testing phases iteratively [9].

It follows an Iterative life cycle model with 5 activities: namely, Planning, design, coding, testing and integration.

- Planning occurs at the beginning of each iteration
- Design, coding, and testing are done incrementally
- Source code is continuously integrated into the main branch, one contribution at the time
- Unit tests for all integrated units. Unlike in case of traditional software development methods regression testing is carried out at all the iterations

B. Working style

The working style is slightly different. It may take sometime for the team members to implement this style.

- Pair programming is followed wherein two programmers work together to develop a particular functionality. This provides a better quality product. There is no concept of peer review.

- Stand up meetings are held to discuss the status of the project. These short meetings are held on everyday basis.

- Collective code ownership: everyone is responsible for the complete software. Anyone can make changes to the program if they find that it will improve the code. Unlike the traditional project management where the project manager would be responsible for the risks, complexity, deadlines etc., in the case of agile software development the team will organize reorganise itself to handle all the pressures of the project.

- The major difference between Extreme programming style planning and incremental delivery, as in case of one of the most accepted "Unified Process" Framework, is the assumption that Architecture and design can be done incrementally. Extreme programming spends very little time on defining the architecture and overall design. It starts directly with the first increment to deliver functionality.

XP undertakes the design in the following manner:

- A *system metaphor* is used which names the classes and the objects in a consistent manner which is agreed upon by all the team members. It also describes the overall architecture.

- It follows the *test first* method where the programmer writes the code for testing the program before writing the program. These tests are written incrementally based on the user requirements. These are the tests the program unit must comply with.

Consider the following class Employee and its corresponding Junit test case :

```
        public class Employee {
        private double Salary;
private String name;

        public Employee() {
                                    name="xyz";
                Salary = 5000;
        }

        public void increment () {
                Salary + = 7000;
        }

        public int getSalary() {
                return Salary;
        }
}
```

The test case in Junit is as follows:

```
import junit.framework.TestCase;

public class MoneyTest extends TestCase {

        public   static   void   main(String[]
args) {

        junit.textui.TestRunner.run(EmployeeT
est.class);
        }

        public EmployeeTest(String arg0) {
                super(arg0);
        }

        protected    void    setUp()    throws
Exception {
                super.setUp();
        }

        protected   void   tearDown()   throws
Exception {
                super.tearDown();
        }

        public void testEmployee() {
        final Employee emp1 = newEmployee();
        assertEquals(5000, emp.getSalary());


        }

}
```

Listing 1, JUnit test for the Employee object

The test case in listing1 tests whether the constructor is invoked and values are initialized correctly.

The code must be written so that it will satisfy these tests.

- *Refactoring reworks* the design to iteratively adapt the changes in the environment and requirements. It follows group design techniques like Whiteboard design sessions, Class-Responsibility-Collaborator sessions (CRC) etc.

Another controversial assumption is that Analysis can be done incrementally XP performs analysis in the following manner:

- User stories, which contain brief description of the requirements, are discussed interactively by a team of programmer and the customer. The requirements are then formulated into acceptance test cases which the software should comply with.

- Planning game is used to do release planning. These are meetings held per iteration. The customer and the programmers participate in the planning game.

The assumption that the product can be delivered to the customer in an iterative manner makes it most beneficial to the customer. The customer can receive and make use of the most important and valuable functionally within a very short period of time. Thus the customers get a sooner return on their investment. The order in which the feature should be added to the software can be decided by the customer. The customer can thus verify whether what is given is what they wanted at a very early stage and can provide feedback to the programmers thus driving the project in the right direction.

C. Is Extreme programming always the best method to follow?

This depends on the project, the team and the customer. If the assumptions of XP hold, there are various benefits to be derived from incremental software delivery. One of them is the early delivery of first incremental model of useful software. Feedback from the customer gives required confidence to the programmer and can help drive the next incremental release. Incremental releases can be scheduled hence there can be predictable delivery. Simple code and minimum documentation saves a lot of time.

III. SCRUM

Another agile software development method is SCRUM.

Scrum and Extreme Programming (XP) are clearly aligned. In fact, if you walked in on a team doing one of these processes you might have hard time quickly deciding whether you had walked in on a Scrum team or an XP team. The differences are often quite subtle, but they are important.

There are four main differences between Scrum and XP [8]:

First, Scrum teams typically work in iterations (called sprints) that are from two weeks to one month long. XP teams typically work in iterations that are one or two weeks long.

Second, in Scrum the teams do not allow changes into their sprints. Once the sprint planning meeting is completed and a commitment has been made to delivering a set of product backlog items, that set of items remains unchanged through the end of the sprint. XP teams are much more flexible within their iterations. As long as the team hasn't started work on a particular feature, a new feature of equivalent size can be swapped into the XP team's iteration in exchange for the unstarted feature.

Third, Extreme Programming teams work in a strict priority order. Features to be developed are prioritized by the customer (Scrum's Product Owner) and the team is required to work on them in that order. By contrast, the Scrum product owner prioritizes the product backlog but the team determines the sequence in which they will develop the backlog items.

Finally, whereas Scrum doesn't prescribe any engineering practices, XP does. XP engineering practices include things like test-driven development, the focus on automated testing, pair programming, simple design, refactoring, and so on.

These are small and often subtle differences between Scrum and XP. They can, however, have a profound impact on the team. A possible approach to maximising the advantages of both approaches is to start with Scrum and then begin using XP engineering practices, such as test-driven development and pair programming.

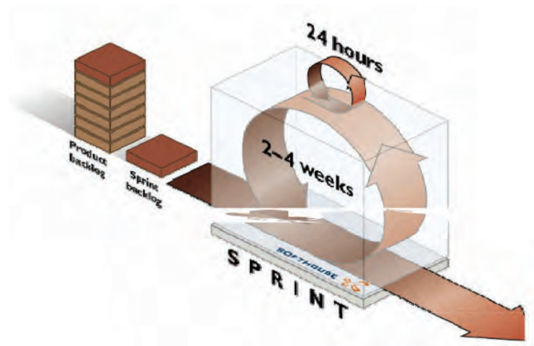


Figure 1: SCRUM [10]

The following graph depicts key differences between the traditional waterfall model and Agile method

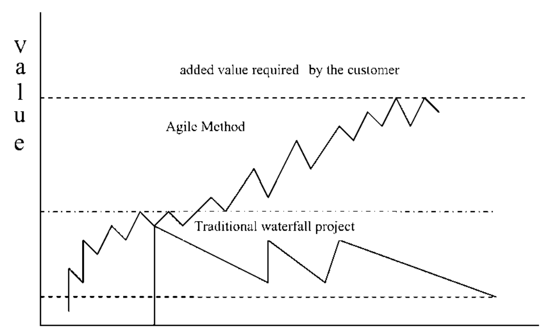


Figure 2: Difference between traditional waterfall model and Agile method.

The graph clearly shows that Agile programming provides added value required by the customer in an incremental fashion while the traditional approach fluctuates between the required value and minimum required value.

IV. DRAWBACKS OF THE APPROACH

If we are adopting agile method for the first time we will face a challenge of changing the mindset of our team member and helping them understand the rules of the new method [5]. A lot of rework may be involved which could have been avoided if a complete requirement analysis is done initially. If a proper order of building the software is not followed, we might end up reworking from the beginning to include another set of requirements. This is especially likely if we need to add certain security or performance features to the software. The complete process may slow down if requirement analysis and design have to be performed for all iterations.

V. CONCLUSION

The Agile software development process will produce incremental release of the software which is very useful to reduce product risks and to delivery a product with customer satisfaction. Delivery of a working model of the product can be delivered at a very early stage. The software will truly be “agile” that is subjected to change until the final delivery. It is always better to have a little planning ahead so as to avoid possible pitfalls. Certain import requirements related to security and performance features can be collected in advance. If one is familiar with the domain, has a stable design and has worked on a similar environment then it is better to follow the waterfall method. If a project is not familiar, has unstable design and involves a lot of risk, it is always advisable to adapt to the Agile software development method. Agile has largely become a synonym for "modern software practices that work," [4].

REFERENCES

[1] Agile Alliance 2007, Available: URL: http:// www.agilejournal.com/ [Accessed 30th Sept 2007]

[2] Agile Manifesto (2007), Available: URL: http:// www.agilemanifesto.org [Accessed 7th Oct 2007].

[3] Beck, K. & Andres, C., *Extreme Programming Explained*. Addison-Wesley, 2004.

[4] Cohn, M.,*“User Stories Applied for Agile Software Development”* Addison-Wesley, 2004.
Construx, http://forums.construx.com/ [Accessed 30th Dec 2007]

[5] Evan, I. “,Agile Delivery at British Telecom” *CSI communications* , July 2007.

[6] IBM http://www.ibm.com/developerworks/java/ library/j-xp042203/ [Accessed 9th June 2007]

[7] Larman, C. (2004). *Agile & Iterative Development.* Addison-Wesley.

[8] Mike (Mike Cohn’s blog) http://blog.mountaingo-atsoftware.com/?p=8

[9] Pascal Van C.” Going round and round and getting nowhere extremely fast?”, CSI Communications, July 2007.

[10] Tamhane, S., IT seminar ‘*Application of IT in heath care’*. Dubai: University of Wollongong , December 2007.

[11] Wells, J. Donovan , *Extreme programming: a gentle introduction*. Available at www.extreme-programming.org [Accessed 9th Oct 2007]

Toward a High Performance Grid Site RO-14-ITIM and the Contribution to WLCG

F. Farcas, G. Popeneciu,
National Institute for Research and Development of Isotopic and Molecular Technologies, Cluj-Napoca, 400293, Romania, email: Felix.Farcas@itim-cj.ro, Gabriel.Popeneciu@cern.ch

Abstract

Grid Technology is the next phase in computer infrastructure that will fundamentally change the way we think about and use computing power. RO-14-ITIM is the Grid site of the National Institute for Research and Development of Isotopic and Molecular Technologies Cluj-Napoca (INCDTIM). In this paper we present the evolution and prospects of the site, his ability to integrate in huge Grid networks, the hierarchical structure as well as his key in research work of the Nord-West part of Romania research field. RO-14-ITIM site support physic domains, HPC and Grid computing for distributed analysis.

I. INTRODUCTION

The Grid computing vision promises to provide the needed platform for a new and more demanding range of applications. To become true, a lot of work, including design and deployment of adequate resource management and information services, need to be overcome. In this context we are presenting the structure of Grid computing that applies at the INCDTIM institute in Cluj-Napoca, Romania.

In 1996 we start working at the installation of the ATLAS LHC CERN Geneva detector with the construction of parts for the Tile Calorimeter. As a next step, studying the forecast in Grid we made the preparation to build up our own Grid site. The work seemed easy but great challenges were to be expected in the next years.

Today we apply and develop the Grid system in many research areas as follows:

(i) The fundamental research in particle physics uses advanced methods in computing and data analysis. Particle physics is the theory of fundamental constituents of matter and the forces by which they interact and ask some of the most basic questions about Nature. Some of these questions have far-reaching implications for our understanding of the origin of the Universe [1].

In 2009 Large Hadron Collider accelerator (LHC) [2] with ALICE, ATLAS, CMS and LHCb experiments, will start taking data. LHC will collide protons at the highest energy $\sqrt{s}$=14TeV and luminosity L=10^{34} $cm^{-2}s^{-1}$ among all the other accelerators and due to these performances, high precision measurements will be possible and the results means new physics.

To fulfil these requirements a high performance distributed computing system is needed.

(ii) During the last years computational simulations based on the atomic description of biological molecules have been resulted in significant advances on the comprehension of biological processes. A molecular system has a great number of conformations due to the number of degrees of freedom in the rotations around chemical bonds, leading to several local minima in the molecular energy hyper-surface. It has been proposed that proteins, among the great number of possible conformations, express their biological function when their structure is close to the conformation of global minimum energy [3].

This type of research involves a large amount of computing power and fills to be a very suitable application for grid technology.

(iii) Over the past few years, a sizable body of experimental data on charge transport in nanoscopic structures has been accumulated.

We face the birth of a whole new technological area: the molecule-based and molecule-controlled electronic device research, often termed “molecular electronics” (ME). The simplest molecular electronic device that can be imagined consists of a molecule connected with two metallic nanoelectrodes. There is now a variety of approaches to form such nanojunctions (NJ) [4] that differs in the types of electrode, materials employed, the way in which the molecule-electrode contacts are established and the number of molecules contacted. Recently the realization of the first molecular memory was reported [5].

The computation of the current-voltage characteristic of a realistic nanodevice needs about 6-10 GB of memory and one week computing time, which bring to the idea of using a Grid environment with high processing power and MPI support.

II. DATA ABOUT ROMANIA GRID SITES

We can speak about Grid sites in Romania after 2001, but officially on 2006 when Romania through National Authority for Scientific Research (ANCS) and The European Organization for Nuclear Research (CERN) signed a “Memorandum of Understanding”, for collaboration in development and exploitation on the Worldwide LHC Computing Grid. [6] From that time, officially we begun to work and develop the Romanian Grid sites.

In Romania we have 2 Grid Consortium, RO-Grid found in 2002 and Tier 2 Federation (ROT2F) found in 2006. The

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_106, © Springer Science+Business Media B.V. 2010

founders of both consortiums are the most important Institutes in Romania which were working in Grid projects years ago, and we can mention the most important: National Institute for Physics and Nuclear Engineering "Horia Hulubei" (IFIN-HH), National Institute for R&D in Informatics ICI Bucharest, University "Politehnica" of Bucharest, Technical University of Cluj-Napoca, Western University of Timisoara and National Institute for R&D of Isotopic and Molecular Technologies (INCDTIM). RO-Grid most important objective is to encourage and facilitate the involvement of other interested and competent national institutions and ROT2F purpose is to provide the computing resources needed to process and analyze the data gathered by the Large Hadron Collider experiments.

Among the first site's in Romania were RO-01-ICI and RO-02-NIPNE and the last one that have a contribution to WLCG is RO-14-ITIM. [7]

III. FIRST STEPS TOWARD INCDTIM GRID SITE

As part of the WLCG Romanian group, and signer of the "Memorandum of Understanding" we started the building of RO-14-ITIM site with 2 servers and a 10 Mbps link in 2005, through the first CEEX project between INCDTIM and IFIN-HH. INCDTIM site was integrated and tested as an extension of the IFIN-HH ATLAS-GRID site. The following tests have been done [8]:

(i) Transmission speed measurements between IFIN-HH and INCDTIM during sustained high speed data traffic, shown in Fig. 1, was obtained using NFS copy. We can see a limited 10 Mbps link, which is not satisfactory for GRID production purposes.

(ii) Entire system functionality tests as a single grid configuration, task distribution and execution real speed depreciation measurements between the remote nodes.

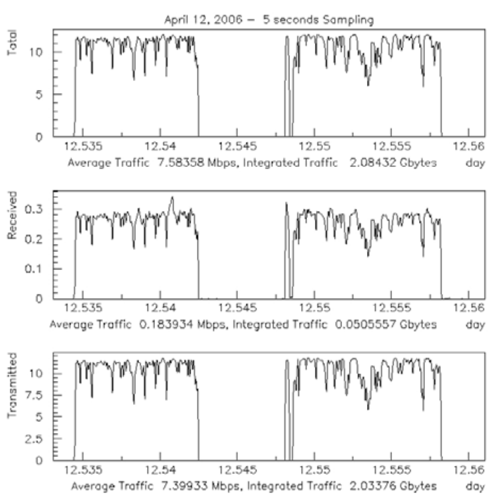


Fig. 1. Traffic measurement between IFIN-HH Bucharest and INCDTIM Cluj on April 2006

For the improvement of our Grid site, this first step where continued through other projects, consisted in purchasing minimal technical equipment.

In parallel we had built a location that meets the highest needs for processing and storing data, which became a type Tier 2 Datacenter [9] with a good security and monitoring system as well as a backup system assured by a 220 kVA generator.

For the final stage we continued to equip our system with a high capacity storage unit and many more computing servers which where included in the INCDTIM grid site, RO-14-ITIM, because the grid system is needed to store and analyze big amount of information.

IV. NOWADAYS GRID SITE RO-14-ITIM

Through the "Memorandum of Understanding", we assumed responsibilities to provide computing power and storage capacity in collaboration with all other WLCG sites from Romania.

The resources that all this system needs are: high speed connectivity, both local and wide area; huge computer throughput, great off-line data storage; human resources to manage local configuration and to implement GRID specific activities.

Nowadays at INCDTIM Cluj there are 3 main directions for processing data:

- ATLAS data processing in WLCG project;
- Numerical simulation of computational physics with application in biophysics and nanostructures;
- Simulation in environment domains.

(i) LCG project - LHC Computing Grid - was launched in 2002 and it has the goal to make a distributed GRID to be used by ALICE, ATLAS, CMS and LHCb experiments. Till now we have in Romania 2 sites that is working for the ALICE project and 3 sites that are working for the ATLAS project. The amount of data that CERN has to process is big enough so that only one site in the world could not work for. The computing resources estimated to be necessary are presented in Tabel.1 [10].

Table 1. ATLAS computing resources estimates

CPU(MS12k)	2007	2008	2009	2010
CERN Tier0	0.91	4.06	4.06	4.61
CERN Tier1	0.49	2.65	4.18	7.00
All tier-1s	4.07	23.97	42.93	72.03
All tier-2s	3.65	19.94	31.77	53.01
Total	9.12	50.61	82.94	138.45

where 1 Athlon 3Ghz ~= 1 kSI2k.

Beside LCG project we intend to use our grid site for other inner projects. Both of them are connected with the numerical simulation of computational physics with direct application in biophysics and nanostructures. As a general rule, the computational techniques used by numerical simulations of computational physics are based on the use of Unix-like operating systems. The codes used are mostly developed in Fortran. A vast majority of these codes are developed under GNU license (such as ABINIT) or academic license (SIESTA, GAMESS etc). The commercial ones are represented by GAUSSIAN or MOLPRO.

The systems under investigation by current applications, require a computing time that can exceeds several weeks [11]. For example, the computation of the current-voltage characteristic of a realistic nanodevice needs about 6-10 GB of memory and one week computing time on the last generation of Opteron processors. Therefore, the development of parallel computational techniques is nowadays in the forefront of development of the methods used by computational physics.

The MPI (Message Passing Interface) has proved to be a very good candidate for this task. The modern developments of the above mentioned codes are using the MPICH2 implementation of MPI in order to reach a good scaling of the computing time with the number of processors.

A. Structure and testbeds

To understand the technical support of INCDTIM pragmatic achievements we point out some theoretical knowledge about network Grid structure.

A Grid model is used, in which a hierarchical structure has been proposed for processing and data storage. This model, generally speaking is based on computing elements, monitoring elements and worknodes as well as a storage element which can be a simple computer or a farm of thousand of hard disks.

A functional testbed needs a scalable infrastructure, adequate middleware and application, sustained by good teams that manage the whole system.

A minimal functional WLCG Grid site has to fulfill some elements as: one computing element, one monitoring element, one testing element, one user interface, one storage element, one site-BDII element and as many as possible worknodes.

It is notable the site administrator role accredited through his valid digital certificate, which is to test, monitor and update the Grid-site having access to all resources within the existing VO in the limits assigned by the ROC staff through the creation process of his site.

B. RO-14-ITIM grid site

RO-14-ITIM Grid Site is an EGEE type, running gLite 3.1 as middleware for the public Network computing elements with Scientific Linux 6.4 and gLite 3.2 for the Worknodes on Scientific Linux 5.3 x86-64.

RO-14-ITIM Grid site is certificated for production and is registered in Grid Operations Center – GOC [12].

The public network consist of one central element (CE), one user interface (UI), one site-BDII server, one storage element (SE) with 80 TB storage capacity a testing element with Nagios and one monitoring element (MON).

The private network contain 54 dual processor quad core servers with 16 GbRam, from which two HP Blade system and one IBM blade system.

The wide area connection is plan to work at 10 Gbps in the near future. INCDTIM local and wide area network operates now at 1 Gbps.

RO-14-ITIM Grid Site configuration is given in Fig. 2.

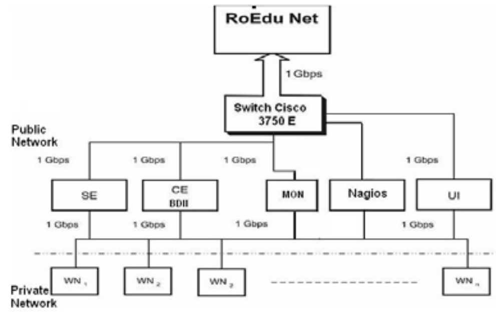


Fig. 2. RO-14-ITIM Site structure

RO-14-TIM hardware configuration, placed in INCDTIM Datacenter, is presented in Fig.3.

Fig. 3. RO-14-ITIM Grid site configuration

The link capacity between IFIN-HH and INCDTIM was monitories in the last 2 years. The results show an increasing link capacity from 200 to 400 MBps after installation in 2009 of a Switch Layer 3 Cisco 350.

Transmission speed measurements during sustained high speed data traffic made in 2008, shown in Fig. 4., carry out a network speed about 200 Mbps.

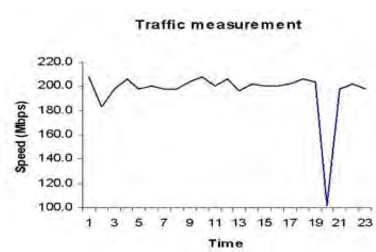


Fig. 4. Traffic measurement on April 2008

After the installation of Switch Layer 3 in 2009, transmission speed measurements, Fig.5, show an increasing to 400-450 Mbps.

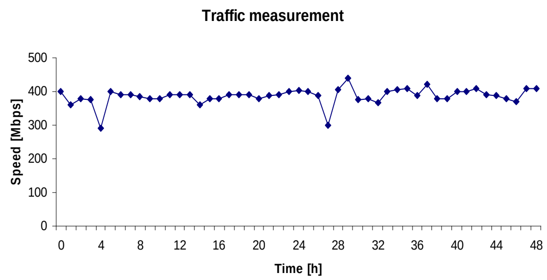


Fig. 5. Traffic measurement on April 2009

After 2009 until now the site was monitorised and compared with other sites. Functionality tests show that the site provides a very good availability and reliability, over 90% for the last 3 months, as shown in Fig. 6.

Site	Phy. CPU	Log. CPU	KSI2K	Availa bility	Relia bility	Unkn own	Availability History Jun-09	Jul-09	Aug-09
RO-01-ICI	12	12	13	93 %	93 %	0 %	100 %	94 %	98 %
RO-02-NIPNE	200	200	300	95 %	95 %	0 %	91 %	89 %	88 %
RO-03-UPB	122	356	677	99 %	99 %	0 %	97 %	90 %	80 %
RO-11-NIPNE	16	32	48	89 %	89 %	0 %	98 %	98 %	82 %
RO-13-ISS	10	22	43	92 %	92 %	0 %	91 %	82 %	53 %
RO-14-ITIM	20	80	328	91 %	91 %	0 %	88 %	97 %	90 %

Fig. 6. RO-14-ITIM availability last 3 month

The statistics for the last 4 month are promising, there where over 500 job done, as shown in the Fig. 7.

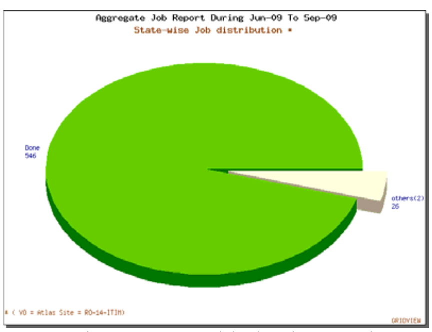


Fig. 7. 500 ATLAS jobs done last 4 month

In comparison with other RO-LCG sites that have a great contribution to WLCG, RO-14-ITIM data are presented in Fig.8.

As a result of this work in August 2009 the site RO-14-ITIM was accepted in the RO-LCG community, most of all because of site great availability.

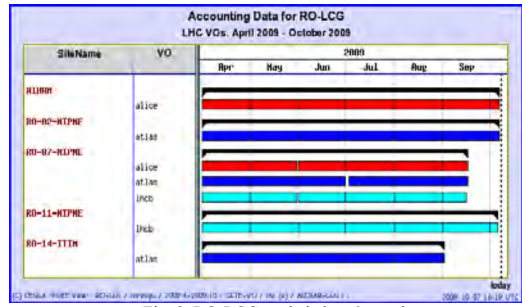


Fig. 8. RO-LCG statistic last 6 month

Fig. 9 shows INCDTIM Datacenter which hosts RO-14-ITIM Grid site (left side of the picture).

Fig. 9. INCDTIM Datacenter

V. FUTURE WORK INTO INCDTIM GRID SITE

Since 2007, when the Datacenter of INCDTIM was finished, National Institute for Research and Development of Isotopic and Molecular Technologies had started for new more environmental friendly resources and power saving computers and storage elements. As a result of this futuristic concept big steps were done in writing and gaining projects. So, through the International project "POS-CCE, Improve of the capacity and reliability of INCDTIM GRID center for its integration into international Grid networks, 2009 – 2010" the institute would be able to improve the storage capacity and processing power of RO-14-ITIM site.

In parallel, a second target is to improve our internal and external bandwidth through the acquisition of a Cisco 6500-E router which would improve the network link to 5 - 10 Gbps with other Grid sites.

To have a reliable Datacenter, with an increased efficiency, we are looking for a water cooling solution, which will also minimize the inside noise.

Reaching these objectives, the side RO-14-ITIM will be able to satisfy both the WLCG specifications and other computing power demands.

VI. CONCLUSIONS

RO-14-ITIM site was designed and developed along almost 5 years, from a basic structure, and with a continuous effort, to a competitive structure, well integrated in TIER-2 Federation.

The RO-14-ITIM contribution to WLCG with his 18.567 CPU time [13] which represents 0.45% of the whole time dedicated by all 5 sites to the CERN project and 3.55% for the ATLAS project.

We interact with new partners in international and national programs and projects, and this give us the opportunity to develop our specialists and infrastructure of grid system, to reach our mission: INCDTIM Datacenter, to be the most important Grid centre in Transylvania capable to sustain the Nord – West research and economical area of Romania.

Acknowledgments

This work is supported by the Romanian Research and Development Agency through EU12/2009, EU15/2009, POS-CEE Ingrid/2008 projects.

References

[1] ATLAS Collaboration, Exploring the Mystery of Matter, Papadakis Publisher, UK, 2008.
[2] LHC Homepage, http://cern.ch/lhc-new-homepage/.
[3] S. P. Brown, S. W. Muchmore, *J. Chem. Inf. Model*., 46, 999 (2006).
[4] A. Salomon et all, Comparison of Electronic Transport Measurements on Organic Molecules, Adv. Mater. 15(22), 1881-1890 (2003).
[5] J.E. Green, et all, Heath: A 160-kilobit molecular electronic memory patterned at 1011 bits per square centimeter, Nature 445(7126), 414 (2007).
[6] Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid, CERN-C-RRB-2005-01, November 2005.
[7] F. Farcas, I. Szabo, A. Bulinschi, L. Miclea, A. Bot , World Wide Romania Grid Database, 3rd International Workshop on High Performance Grid Middleware, 17th International conference on control system and computer science, 26 – 29 May 2009, Proceedings, CSCS – 17 Vol 2, pp 461-464, Copyright © 2009, Editura Politehnica Press
[8] C. Alexa, V. Boldea, M. Ciubancan, Ş. Constantinescu, S. Diţă, D. Pantea, T. Preda, M. Jalobeanu, G. Popeneciu, IFIN-ITIM Distributed Grid System, ISPDC 2006, Proceedings 5th International Symposium on Parallel and Distributed Computing, Timisoara Romania, July 6-9, 2006, pp 215-219, IEEE Copyright 2006.
[9] TIA – 942, Data center standards overview.
[10] J. Knobloch and L. Robertson, LHC Computing Grid, Technical Design Report, CERN 2006.
[11] F. Farcas, G. Popeneciu, A. Bende, C. Morari, S. Belov, L. Miclea, ITIM Distributed Grid System applied in high energy, biomolecular and nanotehnology physics, IEEE International Conference on Automation, quality and testing, robotics (AQTR) May 22-25 2008, theta 16th edition, vol III, Proceedings pp 343-346 2008.
[12] https://goc.gridops.org/site/list?id=1655004
[13] https://www3.egee.cesga.es/gridsite/accounting/CESGA/tier2_view.html

A MFI4OR-based Approach of Semantic Annotation for Web Service

Cheng Zeng
[1]State Key Laboratory of Software Engineering
Wuhan University
Wuhan, P.R. China
[2]College of Mathematics & Computer Science
Hubei University
Wuhan, P.R. China
E-mail: zengcheng8@gmail.com

Keqing He, Bing Li, Chong Wang, Yangfan He
State Key Laboratory of Software Engineering
Wuhan University
Wuhan, P.R. China
E-mail: hekeqing@public.wh.hb.cn

***Abstract*—Compare to the traditional annotation technology for web service, a novel approach for the semantic annotation to web service based on MFI4OR is proposed in the paper. This proposal permits multi-ontology based annotation to web service in the same domain which will lead to the flexible annotation, and then the flexible publication and discovery of web service in Web Service Registry (WSR). The result of experiment illustrates the efficiency of web service discovery and the degree of satisfaction to users in WSR is highly improved.**

Keywords-Semantic Annotation, Web Service, Domain Ontology, MFI4OR

I. INTRODUCTION

With XML based standards like UDDI, WSDL and SOAP, web services play a key role in universal connectivity and interoperability of applications and services [1]. With the growing popularity of Web services, there arise issues of finding relevant services, especially with the possibility of the existence of thousands of Web services. Web Service Registry (WSR), which acts as an information mediation role in the triangular SOA, is to solve this problem. In WSR, service providers can publish service description information more effectively and service consumers can discover appropriate services more accurately and easily. However, currently web services are described as WSDL descriptions, which provide operational information such as Inputs and Outputs etc. It is a type of syntactic level description and its interface parameters do not have meaningful names, which obstructs service discovery and composition. For example, one service A has an output item which use *specialty* as name to describe what field an expert engages in, while another service B may use *research field* for describing such an item. As we all know, *specialty* and *research field* represent the same information item, but they are different at syntactic level, and same at semantic level. Therefore, the key to solve the limitation of keyword-based searches is to add semantic annotations to Web service specifications [2].

Domain ontology is used to add semantic information on WSDL interface parameters to overcome the limitations of a syntax-based search [3]. The discovery of web services is enriched with the adoption of semantic annotation to improve the flexibility of the WSR and help users with solutions (services) that are close enough to what they would have liked to get (even if they do not fully match their expectations). Therefore, how to annotate these items becomes a key point in the issue of semantic annotation. In recent years, a large number of research based on semantic techniques have been proposed, such as the efforts around OWL for Services (OWL-S) [8], the Web Services Modeling Ontology (WSMO) [9], WSDL-S [10] and METEOR-S [12].

In this paper, a novel approach for the semantic annotation for web service based on ISO SC32 19763-3: MFI4OR (Metamodel Framework for Interoperability: Metamodel for Ontology Registration) [4] is proposed, which will make semantic annotation more accurate and efficient. The remainder of the paper is organized as follows: Section 2 presents the traditional model of semantic annotation. In Section 3, we provide some improved methods of semantic annotation by use of domain ontology. Section 4 proposes the MFI4OR-based approach of semantic annotation. The related work is discussed in Section 5, followed by the conclusion and future work in Section 6.

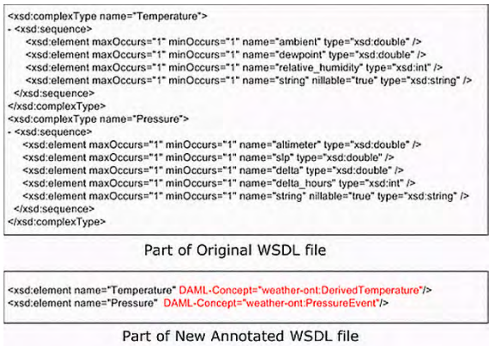
II. TRADITIONAL MODEL OF SEMANTIC ANNOTATION

WSR is supported by domain ontology for semantic publication and discovery of the Web services. The domain ontology is created from concepts and terminologies that are likely to be used by Web services in a particular domain.

For example, in the electronic commerce domain, the domain ontology can consist of concepts from the ebXml Core Component Dictionary. Another example could be domain ontology for the Traffic domain which may include concepts like *Car, CarPrice, ArrivalCity* and *DepartureCity*. As a result, before the semantic annotation of web service, we must choose a concrete domain ontology. After that, we can manually use the mapping between these items in WSDL file (mainly refers to WSDL interface parameters of Web Service such as Inputs,

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_107,

Outputs and Operations) and the selected domain ontology to capture the meaning implied by the Web service provider in that domain. Finally, the mappings are written to the WSDL file as annotations. The annotated WSDL file along with the original WSDL file is shown in figure 1. However the manual process of annotation can not assure the accuracy of semantic annotation. Section 3 is ready to overcome this problem.



```
<xsd:complexType name="Temperature">
- <xsd:sequence>
    <xsd:element maxOccurs="1" minOccurs="1" name="ambient" type="xsd:double" />
    <xsd:element maxOccurs="1" minOccurs="1" name="dewpoint" type="xsd:double" />
    <xsd:element maxOccurs="1" minOccurs="1" name="relative_humidity" type="xsd:int" />
    <xsd:element maxOccurs="1" minOccurs="1" name="string" nillable="true" type="xsd:string" />
  </xsd:sequence>
</xsd:complexType>
<xsd:complexType name="Pressure">
- <xsd:sequence>
    <xsd:element maxOccurs="1" minOccurs="1" name="altimeter" type="xsd:double" />
    <xsd:element maxOccurs="1" minOccurs="1" name="slp" type="xsd:double" />
    <xsd:element maxOccurs="1" minOccurs="1" name="delta" type="xsd:double" />
    <xsd:element maxOccurs="1" minOccurs="1" name="delta_hours" type="xsd:int" />
    <xsd:element maxOccurs="1" minOccurs="1" name="string" nillable="true" type="xsd:string" />
  </xsd:sequence>
</xsd:complexType>
```

Part of Original WSDL file

```
<xsd:element name="Temperature" DAML-Concept="weather-ont:DerivedTemperature"/>
<xsd:element name="Pressure"  DAML-Concept="weather-ont:PressureEvent"/>
```

Part of New Annotated WSDL file

Figure 1. Fragment of original and annotated WSDL files

III. Improved Model of Semantic Annotation

In this section, a semi-automatically mapping algorithm is used in the process of semantic annotation which will improve the accuracy of semantic annotation. Figure 2 shows the improved annotation process of mapping WSDL concepts to the nodes in Domain during semantic annotation. Different from traditional model, we have developed an algorithm to semi-automatically map each individual concept in the WSDL description (i.e. Inputs, Outputs, and Operations) to an ontological concept.

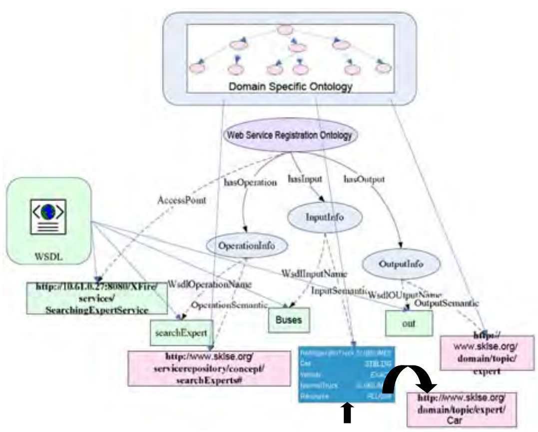


Figure 2. Improved Model of Semantic Annotation

To achieve this goal, the algorithm relies on two kinds of ontologies: the domain ontology which is discussed in section 1 and the general-purpose ontology. The first includes terms related to a given application domain. We assume that this ontology can be built by domain experts who analyze the terms included in the Web services published in the registry. The latter includes all the possible terms (at this stage we use Wordnet).

We assume that programmers of Web Services have a general understanding about the domain which Web Services belong to, such as the *Electronic Commerce* domain or the *Traffic* domain. Therefore, when defining the names of operations and parameters in Web Service, they may use the concept in the domain, which can avoid the naming at will. When annotating the names of operations and parameters, our algorithm will realize the following two steps: (1) Making a match between these names and the concept items in the domain ontology, and producing the candidate concepts list from the domain ontology to annotate. (2) Sorting the candidate concepts by the priority of (Exact, Plug-in, Subsume, Fail) according to the algorithm introduced in the [7]. The sorted candidate concepts list is shown in Figure 2. Then the publisher can select the most satisfying concept from candidate concepts to annotate the names used in Web Service.

However, it is possible that the names of operations and parameters may not exist in the domain ontology, or reflect the naming convention usually adopted by programmers: *UserName* or *currencyExhange* are more frequent than the simple names directly included in the ontology. For this reason, if the names are composite words, we will tokenize the word and return the similar word to replace it; if the names are not directly included in the ontology, our algorithm will make use of the general-purpose ontology (i.e. WordNet) to build the synset list of the missing name. And then each word in the synset will be made matching in the same way introduced above. The sorted candidate concepts list is also created for the provider to annotate.

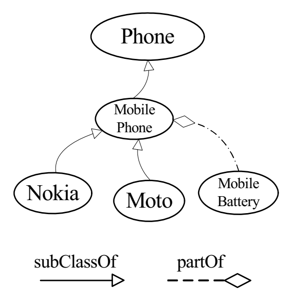


Figure 3. Fragment of Ontology

In addition, our algorithm considers the other match-making relationship: *partOf* and *siblingOf*. In [7], the algorithm only proposes the four degree of matching: Exact, Plug-in, Subsume and Fail, which has some disadvantages in matchmaking. For example, Figure 3 shows an fragement of phone ontology. If one service has an output named as *Nokia*, the concept *Mobile phone* and *phone* will be selected as candidate concepts according to the algorithm in [7]. In fact, the concept *Moto* should also be considered as the candidate concept because it has a *siblingOf* relationship with the concept *Nokia*. For another case, if one service has an output which uses *Mobile battery* as name, there is no matching concept using the algorithm in [7]. Actually, the concept *Mobile Phone*, which has a *partOf* relationship with *Mobile battery*, has a possible opportunity to be the candidate concept. Our algorithm has considered these possible matching relationships between the signatures of Web Service and the concept items in the domain ontology when annotating, which ensures to achieve the satisfying annotation at most.

Whatever traditional model or improved model mentioned in section 2 and 3, they are both focus on the single domain ontology to annotate. In section 4, the method of multi-ontology based annotation to web service is proposed which will lead to the flexible annotation in WSR.

IV. MFI4OR-BASED MODEL OF SEMANTIC ANNOTATION

A. *What is MFI4OR?*

Ontology provides a basic technology to support the semantic interoperability of different systems over a distributed and heterogeneous environment. If the ontologies have the semantic relation, they could understand each other to some extent, which is the precondition of semantic interoperability. MFI4OR is mainly used for providing a common framework to register ontology and its evolution information for semantic interoperation between ontologies or web-based applications based on them. It provides the solutions by defining "Reference Ontology" (RO) and "Local Ontology" (LO) which are both registered by a common structure formed as *Ontology-Ontology Component-Ontology Atomic Construct* [4]. The Ontology Components defined in RO can be reused by the other ontologies, while that defined by LO can only for its own definition purpose. RO is always constructed by domain experts, while LO is evolved from RO. The LOs and RO can interoperate with each other, which provide a solid foundation for interoperability between the different ontologies. Besides RO and LO, the ontology evolution rules between them are the core part in MFI4OR which can help record the detailed changes when LOs evolve from RO. These rules can ensure that the different ontologies used to annotate web service will not hamper the interoperability among them.

General speaking, there always exist three rules: *SameAs* Rule, *Enhancement* Rule, and *Contraction* Rule [5, 6]. *SameAs* Rule can set up the equivalent mapping between instances of Ontology, Ontology Component or Ontology Atomic Construct. *Enhancement* Rule is special for adding some new elements to the original ontology when versioning and evolving. *Contraction* Rule is often adopted in ontology evolution management to delete concepts, attributes, properties, constraints and so on. Theoretically, the three basic rules can cover all the possible modifications on RO and LO because any complex evolution can be decomposed, viewed as a sequence consisting of these three operations.

B. *MFI4OR-based Model of Semantic Annotation*

As mentioned in section 2 and section 3, the process of semantic annotation to web service is based on the single domain ontology. However there also exist some cases that different ontologies in the same domain are used to annotate web service, which will lead to the issue of the interoperability of multi-ontology in the same Domain.

As we all know, Domain Ontology is always constructed and standardized by Domain Experts which use common approved concepts and relationship among them, but sometimes service providers maybe define their own concepts to semantically annotate their web services for meeting their requirements. We regard the domain ontology defined by experts as **Domain Reference Ontology** (DRO) and the domain ontology defined by users as **Domain Local Ontology** (DLO). However the process of defining the new concepts in DLO is not at will, but evolved from DRO according to three ontology evolution rules (*SameAs* Rule, *Enhancement* Rule, and *Contraction* Rule) defined in MFI4OR. That is to say, DRO and DLO are not isolated but interrelated to each other based on these evolution rules. Figure 4 shows that DRO and DLO can both be adopted to add semantic information to Web Service. It can extend the scope of semantic annotation in the Domain, which will enhance the flexibility during the process of service publication and discovery.

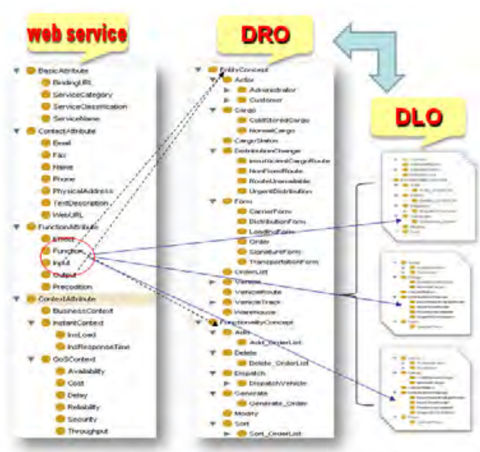


Figure.4. Multi-Ontology annotation Based on MFI4OR

As mentioned above, DLO is evolved from DRO, and the evolution information between DRO and DLO should be registered and recorded, which will improve the interoperability between them in WSR. Although there exist three evolution rules between DRO and DLO, we are ready to adopt *SameAs* rule because this rule is the part of the first version of MFI4OR which has already become International Standard since 2008, and the other two evolution rules will be in the second version of MFI4OR in the future.

C. Realization of the MFI4OR-based Model

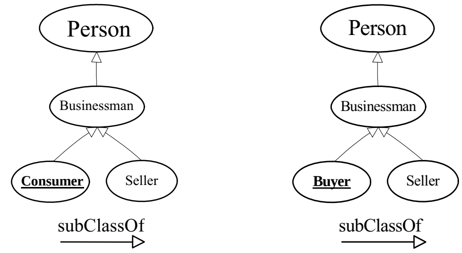
Figure 5 shows the User Interface (UI) of annotation with DRO or DLO. Once service provider selects the Domain that web service belongs to, such as *logistics* or *transport*, DRO and DLO will be imported to this platform for the use of annotation to web service. The evolution information between DRO and DLO will lead to the interoperability between them.

Figure.5. UI of annotation with DRO or DLO

For example, there is a fragment of DRO shown in Figure 6 (a). When evolving from DRO to DLO, Service Providers can reuse the concepts of *Person*, *businessman* and *seller* in DRO, and more important can define the new concept of *buyer* for their requirement shown in Figure 6 (b). Meanwhile we adopt *SameAs* rule to record the evolution information between ***consumer*** in DRO and ***buyer*** in DLO shown in table 1. It shows that DLO is not separated from DRO but evolved from DRO through the evolution information. If service provider use the concepts in DLO to annotate web service, when querying the web service whose output is *buyer*, the returned candidate list of web service will include the web service whose output is *consumer* according to evolution information besides the web services whose output is *businessman* or *person* according to traditional semantic model. That is to say, this model of semantic annotation will return the more web services to satisfy the users.

Table 1. Evolution information of DLO

Evolution information	
attribute of LO	*Value*
Name	*Buyer*
URI	*URI_ buyer*
modelType	*OWL*
beforeChange	*Consumer*
afterChange	*Buyer*
transformationRule	*SameAs Rule*
consistentChange	*True*



(a) DRO (b) DLO

Fig.6. Fragment of DRO and DLO

D. Evaluation

To evaluate our approach, we take precision, the average time and the average number of web services as the evaluation metrics. Precision is defined as the ratio of correct web services retrieved to the total number of result retrieved. The performance is measured on a workstation of 1GB RAM, 2.6GHz CPU with Microsoft Windows XP.

Evaluation is performed on TMSA (Traditional Model of Semantic Annotation), IMSA (Improved Model of Semantic Annotation), and MMSA (MFI4OR-based Model of Semantic Annotation). Figure 7 shows the relationship of retrieving time and the number of web service in different models of semantic annotation. The average retrieving time is increased with the increasing number of web services. Meanwhile in MMSA, the average time of retrieving web service from is also increased compared to two other models because of the time-consuming on evolution operation between DRO and DLO in the same domain. Although the average retrieving time is spent more in MMSA, the high precision can be achieved because it can return more satisfied web services to be chosen during retrieving time. See Figure 8.

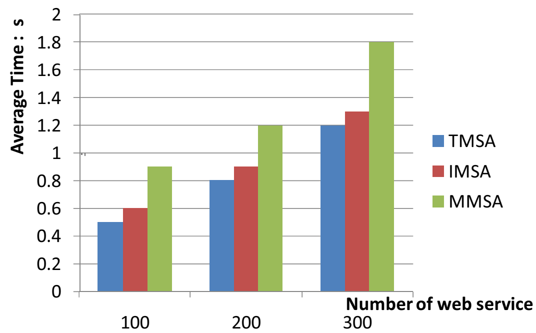


Figure 7. The relationship of retrieving time and the number of web service in TMSA, IMSA, MMSA

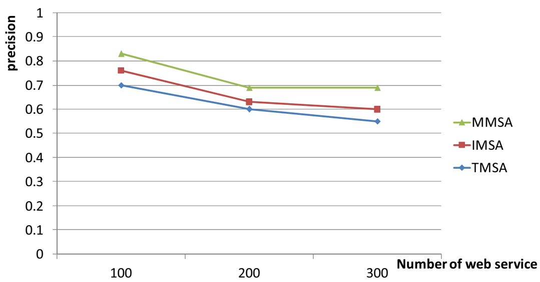


Figure 8. The relationship of precision and the number of web service in TMSA, IMSA, MMSA

V. Related Works

Our work presents an approach for adding semantics to Web services based on MFI4OR. In fact, much work has been done on semantic annotation to Web services. We discuss two representatives of them. A classification method is described in [11]. It uses machine learning algorithms to help users simplify the search for relevant ontology. Unfortunately, the method considers just names of WSDL concepts, ignoring the structure of them. Moreover, it uses vocabularies, but does not use ontology which is more descriptive and capture domain more accurately for classification. MWSAF [12] uses a media structure called SchemaGraph to facilitate the matching between XML schema and ontology to find the relevant ontology. However, since the matching results are not good if WSDL files do not have good structure or the ontology becomes much comprehensive, it is also not suitable for our context. Moreover, [13] proposes a template-based markup tool, which can improve annotation accuracy significantly. But as used for semantic web content annotation, the template is over complex, and has a poor reusability considering the various web contents.

The common factor in these approaches is only relating signatures of Web service to SINGLE domain ontology, but our approach can use DIFFERENT ontologies in the same domain, which are not isolated but interrelated to each other based on ontology evolution rules defined in MFI4OR, to annotate web service, which will extend the scope of semantic annotation in the Domain and greatly enhance the flexibility during the process of service publication and discovery.

VI. Conclusion and Future Work

At present, few academic models of semantic annotation extend beyond simple subsumption relationship based input and output matching in single domain ontology. In this paper, in IMSA we have considered all possible matching relationships between the signatures of Web Service and the concept items in the domain ontology when annotating. More important, in MMSA we are not limited to single domain ontology to annotate. Different ontologies in the same domain based on MFI4OR can be adopted when annotation which will improve the efficiency of semantic annotation.

As future work, we plan to expand the Ontology Evolution Rules which make the evolution information to be recorded comprehensively and clearly which will achieve more satisfied annotation results to fulfill user's requirements.

Acknowledgments

This work is supported by National Basic Research Program (973) of China under grant No. 2007CB310801, National High Technology Research and Development Program (863) of China under grant No. 2006AA04Z156, National Natural Science Foundation of China under grant No. 60873083,60803025, 60970017 and 60903034, Natural Science Foundation of Hubei Province for Distinguished Young Scholars under grant No. 2008CDB351, Natural Science foundation of Hubei Province under grant No. 2008ABA379, Research Fund for the Doctoral Program of Higher Education of China under grant No. 20070486065 and 20090141120022, China Postdoctoral Science Foundation under grant No. 20090460107, and the Research Project funded by Wuhan University under grant No. 6082005 and 6082007, Scientific and Technological research projects of Hubei Province Education Department under Grant No.D20091012.

References

[1] R. Chinnici, J. Moreau, A. Ryman, and S. Weerawarana, Eds., "Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language", W3C Candidate Recommendation 27 March 2006; http://www.w3.org/TR/2006/CRwsdl2020060327

[2] K. Verma and A Sheth, "Semantically Annotating a Web Service," IEEE Internet Computing 11 (2),March–April 2007, pp. 83–85

[3] Jian Wang, Keqing He, Wei Liu, Rong Peng, Bing Li, "A Service-Oriented Domain Modeling Framework for Networked Software", Dynamics of Continuous Discrete and Impulsive Systems - Series B, Special Issue on Software Engineering and Complex Networks, December 2007, pp. 246-252

[4] ISO/IEC 19763-3:2007. Keqing He. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38637

[5] Chong Wang, Keqing He, Yangfan He. MFI4Onto: Towards Ontology Registration on the Semantic Web [A]. In Proceedings of 6th International Conference on Computer and Information Technology (CIT 2006) [C], Seoul, 2006, pp.40-46

[6] Cheng Zeng, Keqing He, Zhitao Tao. Towards Improving web service Registry and Repository Model Through Ontology-Based Semantic Interoperability, GCC2008, 2008, pp.747-752

[7] Paolucci, M, Kawamura, T, Payne, T. R, and Sycara,K, Semantic Matching of Web Services Capabilities, In Proceedings of the 1st International Semantic Web Conference (ISWC2002).

[8] D. Martin, M. Burstein, J. Hobbs,OWL-S: Semantic Markup for Web Services", W3C Member Submission, Nov.2004; http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/.

[9] H. Lausen, A. Polleres, and D. Roman, "Web Service Modeling Ontology (WSMO)", W3C Member Submission, 2005;http://www.w3.org/Submission/WSMO/.

[10] R. Akkiraju, J. Farrell, J. Miller, M. Nagarajan, M. T. Schmidt, A. Sheth and K.Verma, "Web Service Semantics - WSDL-S", W3C

Member Submission, 7 November 2005; http://www.w3.org/Submission/WSDL-S/.

[11] A. Hess and N. Kushmerick. Automatically attaching semantic metadata to Web services. Proceedings of the 2nd International Semantic Web Conference (ISWC 2003).

[12] Patil A, Oundhakar S, Sheth A, Verma K. METEOR-S Web service annotation framework. In Proceedings of the 13th Int'l World Wide WebConference, New York, ACM Press, 2004, pp.553-562.

[13] Brian Kettler, James Starz. A Template-Based Markup Tool for Semantic Web Content. Proceedings of ISWC2005, LNCS 3729, 2005.

Ethernet Node for Serial Communications

A.S. Popa[1], M. Popa[2], T. Slavici[3], D. Darvasi[4]
[1,2,3]"Politehnica" University from Timisoara, Romania
[4]"Ioan Slavici" University from Timisoara, Romania

Abstract—**Several protocols exist in serial communications, such as SPI, USB, RS232, RS485, CAN etc. In order to obtain flexible and fast communication, serial communication nodes were developed. They are generally based on microcontrollers and can be seen as protocol converters. Their capabilities were enriched by adding the remote command possibility, through an Ethernet interface. This paper presents a serial communication node based on the powerful 32 bit MPC5553 microcontroller. It implements a bridge between two SPI, two RS485 and one Ethernet interface. It is used for collecting data from sensors and sending them over the Internet to a testing station.**

I. Introduction

Several protocols exist in serial communications, such as I2C, SPI, USB, RS232, RS485, LIN, CAN etc. They are more or less appropriate for certain applications. For example, the CAN protocol is very used in automotive applications, due to its reliability, RS232 protocol is used in industrial and PC based applications, due to its simplicity and low cost and so on.

In order to obtain flexible and fast communication, serial communication nodes were developed. They are generally based on microcontrollers and can be seen as protocol converters. Their capabilities were enriched by adding the remote command possibility, through an Ethernet interface.

Such microcontroller based serial nodes exist, with different performances. In reference [1] a node for wireless sensor networks is presented. It is based on an Atmel ATmega 128L microcontroller with two UART, one SPI and one I2C interfaces and with input ports for sensors. Reference [2] describes a network node based on an Atmel ATmega 128L microcontroller, which has an I2C, two UART and one Bluetooth interfaces. In reference [4] a powerful node is presented. It is based on two microcontrollers, an 8 bit Atmel ATmega 128L and a 16/32 bit Atmel AT91FR4081 ARM THUMB and includes I2C, RS232, RS485, SPI and RF interfaces.

This paper presents a serial communication node based on the powerful 32 bit MPC5553 microcontroller. It implements a bridge between two SPI, two RS485 and one Ethernet interface. It is used for collecting data from sensors and sending them over the Internet to a testing station. The next section shortly describes the SPI and RS485 protocols, the third section presents the new solution and the fourth section outlines the conclusions and future development directions.

II. The SPI and RS485 Serial Protocols

The SPI protocol describes the communication via a full duplex, master – slave interface. The interface lines are:

- SCLK: common clock for all the devices: it is an output for a master device and an input for a slave device;
- MISO and MOSI: data lines; for a master device MISO is a data input and MOSI is a data output and for a slave device the roles of the two lines are inversed; all the MOSI lines are connected together and all the MISO lines are connected together; there is only one master which will command the SCLK line and will generate data on its MOSI line to the MOSI inputs of the slaves; only the selected slave will generate data on its MISO output to the MISO input of the master;
- /SS: the significance of this line differs; if the device is a slave one, then /SS is an input for its validation and the master must generate so many /SS lines how many slaves are therefore for a system with n slaves there are n + 3 lines; if the device is a master one, /SS may be output or input, its significance being software selected; if it is an output it will validate a slave and if it is an input it can be used for error detection or to receive an external request from other device for the master position.

RS485 is a protocol for serial asynchronous communication between more devices directly connected to the same lines. The connection is single master many slaves type. The logical levels are represented by differential voltages which increase the number of lines but the protection to noises too leading to a maximum distance between the devices of 1200 m.

III. The Proposed Solution

A. The Hardware

Fig. 1. presents the bloc diagram of the proposed serial node. It is based on the powerful MPC5553 microcontroller, [3], based on the PowerPC Book E architecture. It collects data through two internal SPI interfaces. Each such interface is connected to SPI lines and to RS485 lines. Both interfaces were connected in order to ensure compatibility with newer sensors which transfer data through the SPI bus, and with older sensors which transfer data through the RS485 bus. A differential driver was foreseen on the SPI lines because of the high distance between the sensors and the microcontroller. The RS485 lines are connected to the SPI lines of the microcontroller through a RS485 – SPI adapter. The Ethernet interface was implemented with a LAN8187i transceiver connected to the internal MII interface. The MII interface ensures the connection to the internal specific FEC (Fast Ethernet Controller) peripheral. FEC corresponds to the MAC layer and requires an external transceiver for being connected to the physical interface. The CAN interface will be used to connect the module to other blocks from the vehicle, connected to the CAN bus and the SPI3 interface will be used for writing parameters into the EEPROM internal memory through an USB interface.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_108,

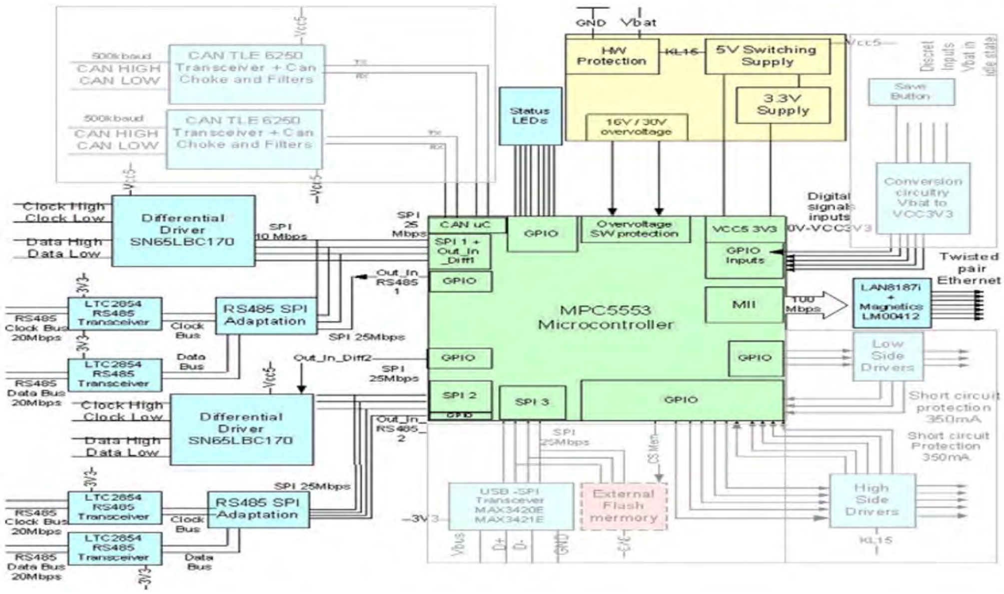


Fig.1. The bloc diagram of the serial node

The MPC5500 microcontroller has the e200z6 core which implements a version of the PowerPC Book E architecture. This is appropriate for high performance embedded systems applications. The e200z6 core integrates an arithmetic execution unit, a branch control unit, units for fetching instructions and for reading/ writing data and a multiport register able to support up to three read operations or two write operations in the same clock period. Most of the instructions are executed in a single clock period. E200z6 is a scalar core with 64 general purpose registers. Floating point instructions are not supported by the hardware but are emulated, through the traps. It exists also a signal processing extension for supporting floating point operations in real time. Other features of the e200z6 core are: powerful interrupts/ exceptions handling unit, memory management unit with full associative 32 inputs TLB, 2 – way set – associative 8 Kbytes cache, delay and watchdog functions and low consume modes.

The PWM signals are obtained by program. Internal counters are used and their content (C) is continuously compared with programmed values (PW). If $C < PW$, 0 logic will be sent on the PWM output and if $C \geq PW$, 1 logic will be sent on the PWM output.

B. *The Software*

The application was conceived in Visual C++ based on MFC (Microsoft Foundation Class) library which permits easy implementation of the Windows based programs. The application has a graphical interface with dialog windows specific to the Microsoft Visual C++ environment.

The SPI driver has the following roles:

• initialization of the attributes of the transfer: speed, number of bits/frame, polarity and phase of the clock signal, initial mode,

• data reception: the bytes received from the SPI interface are buffered until the reception of the whole multiframe and

• data transmission: the data bytes are sent over the SPI interface.

The Ethernet driver has the following roles:

• driver initialization: establishment of a connection to the Ethernet and establishment of the MAC address,

• Ethernet frames reception: frames are buffered,

• Ethernet frames transmission: an Ethernet frame is sent over the MII interface to the external transceiver which will send it on the Ethernet and the maximum length of an Ethernet frame is 1518 bytes.

The communication protocol between the sensors, the node and the Ethernet achieves the following operations:

• the frames received from the Ethernet are unpacked and the command contained in the data field is sent on the SPI bus;

• the Ethernet frames with a destination address different than the MAC address of the serial node are ignored;

• the frame received from Ethernet is not verified; it is supposed as being correct;

• the bytes collected from the sensors are packed in one or several frames and are sent on the Ethernet;

• the destination address for the frames is broadcast type;

• in case of errors at an Ethernet frame transmission, for example because collisions are detected, a retransmission is tried and after 10 such tries the transfer is aborted and error situation is indicated through a led.

The software was conceived in two layers:

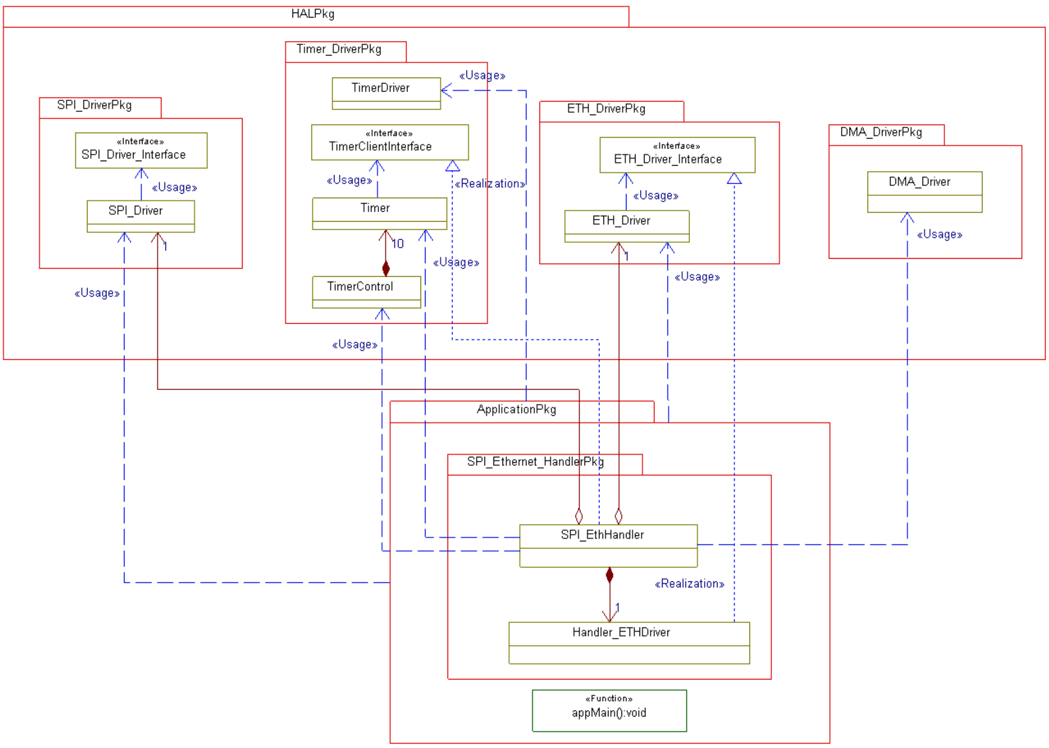


Fig. 2. The hierarchy of packets and classes

• the Hardware Abstraction Layer: it consists of the classes corresponding to the microcontroller interfaces; it contains the following classes: `SPI_Driver`, `ETH_Driver`, `DMA_Driver` and `TimerDriver;`

• the Application Layer: contains the application, more exactly the protocol between the sensors, the serial node and the Ethernet; it contains the `SPI_EthHandler` and `Handler_ETHDriver` classes.

This structure is implemented by a hierarchy of packets and classes. Fig. 2. presents this hierarchy and fig. 3. presents the UML diagram of the software architecture.

Fig. 4. presents the UML diagram of the `SPI_Driver` class. The singleton design pattern was used for implementing this class. The reason is that the microcontroller has three SPI interfaces, therefore the class must be instantiated three times. The `spi_GetInstance` static method returns the instance of the SPI channel specified by the `e_spi_channel` parameter.

The `SPI_Driver` class contains methods for initializing the internal DSPI peripheral, for transmitting and receiving SPI frames and for configuration.

The method `v_Init()` initializes the SPI internal peripheral. The terminals and the attributes of the transfers are configured. The method `T_VOID v_Send(T_UWORD transmitData, T_BOOL b_EOQFlag)` has the role of sending commands and data on the SPI interface.

The `SPI_Driver class` uses the `SPI_Driver_Interface` interface for implementing the callback mechanism. There are two methods `for` enabling, (`b_AttachCallbackHandler(SPI_Driver_Interface*CallbackHandler)),` and disabling, (`v_DetachCallbackHandler()),` this mechanism. It informs the client that the transmission was successfully or that a new SPI frame was received and it was sent to the client. Thus, the client will be informed automatically about the occurrence of an event without the need of continuously interrogating the class `SPI_Driver` about that event. The `SPI_Driver_Interface` specifies the methods `v_RxCompleted(T_UWORD received_data)` and `v_TxCompleted()` for the two events above mentioned.

Fig. 5. presents the UML diagram of the `ETH_Driver` class. It implements the singleton pattern and defines the interface corresponding to the FEC peripheral. It contains methods for initializing the FEC peripheral, the LAN8187i transceiver, for transmitting and receiving Ethernet frames and for other configurations. The return of an instance of the class is done through the `eth_GetInstance()`method.

The method `ub_Init()` configures the pins corresponding to the MII interface and initializes the internal registers through which the origins of the memory zones allotted to the circular lists of the transmission and reception descriptors and the dimension of the reception buffers are specified.

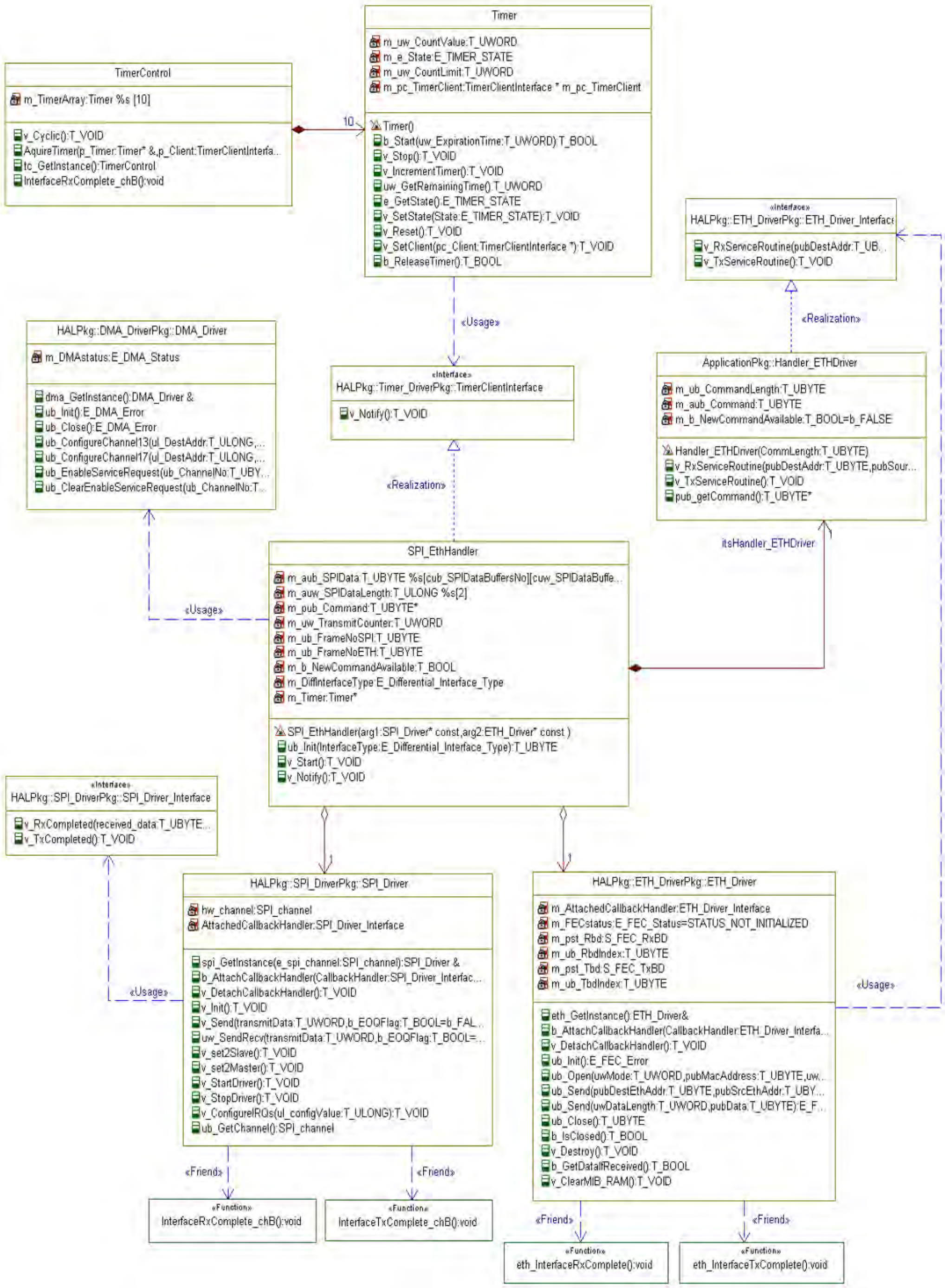


Fig. 3. The UML diagram of the software architecture

The method `ub_Open(T_UWORD uwMode, const T_UBYTE* pubMacAdress, T_UWORD uwPhyType)` achieves the operations necessary to the FEC peripheral to be ready for transmitting and receiving. These operations are: initialization of the descriptor lists, specification of the MAC address, configuration of the control registers for transmitting and receiving and initialization of the PHY external transceiver. For initializing the descriptor lists, two private methods were foreseen, `v_RxBDInit()` and `v_TxBDInit()`, called from the `ub_Open()` method.

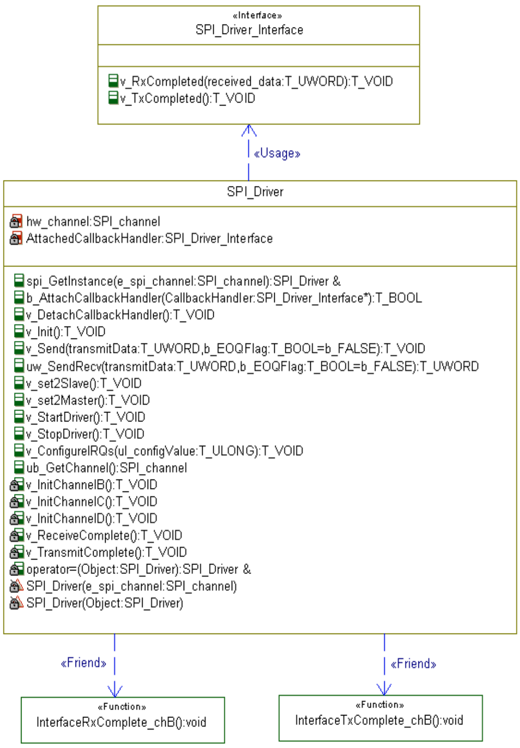


Fig. 4. The UML diagram of the `SPI_Driver` class

The initialization of the PHY transceiver implies the configuration of its internal registers. This is done through reading and writing management frames. The operations are implemented in two dedicated methods: `ub_MiiWrite(T_UBYTE ubPhyAddr, T_UBYTE ubRegAddr, T_UWORD uwData)` and `ub_MiiRead(T_UBYTE ubPhyAddr, T_UBYTE ubRegAddr, T_UWORD* puwRetVal)`. Their parameters specify the address of the transceiver PHY, the address of the register and the value which will be written in, respectively a pointer to the memory location where the read value will be stored. The PHY address is estableshid according to the hardware configuration of certain pins of the circuit.

The parameters of the connection can be by established by configuration or through the auto – negotiation process. This is a mechanism for exchanging information between two interconnected nodes and in which the operation mode with higher performance is selected among the operation modes supported by the two nodes. The auto – negotiation process can be started automatically when the connection beween the two nodes is lost, by setting a specific bit or can be started by software. The configuration information sent by the PHY transceiver to the partner during the auto – negotation process, are saved in specific registers. At the end, one register contains the partner's parameters and an other one contains the connection's parameters. The results of the auto – negotiation operation can be used by the FEC peripheral too, for example it can set the communication type as being half – duplex or full – duplex.

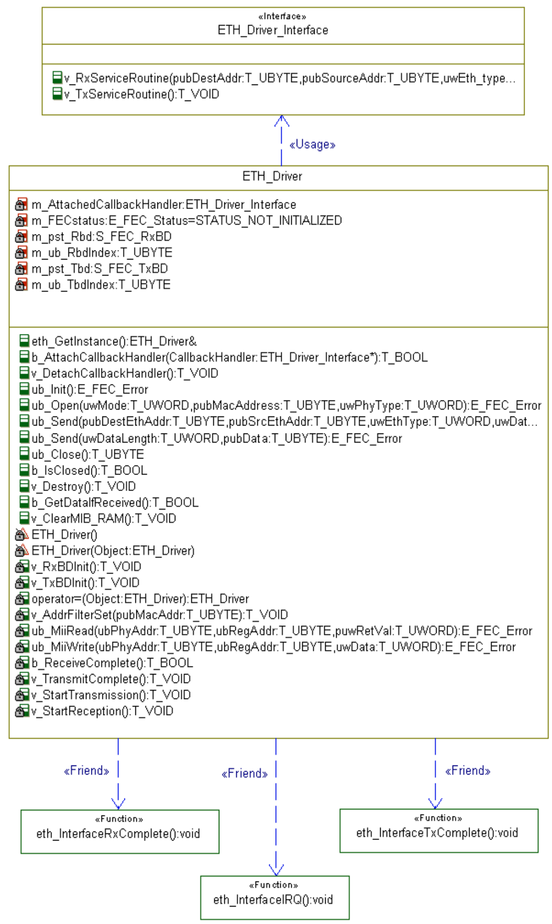


Fig. 5. The UML diagram of the ETH_`Driver` class

In the transmission, a frame will be divided in multiple buffers. For each occupied buffer, a flag will be set. The last buffer containing the frame will have a specific flag activated.

There is no a predetermined relation between a frame and an Ethernet header, as a consequence one can have a descriptor corresponding to the data buffer and an other one corresponding to the buffer containing the Ethernet headers. The two transmission flags, one for the data and the other for the header, must be set in a specific order: first the one corresponding to the data and second the one corresponding to the header. Otherwise, an error will be indicated. There are frames with the same or with different Ethernet headers and different methods will be used for sending them. The transmission is done through DMA channels. Aftre reading a buffer, the flag indicating its state will be reset, thus the software will know which buffers are free for an other incoming frame.

At the reception, the length of the frames is not previously known. It is necessary to set a variable for defining the length of all the reception buffers. A frame may occupy several buffers and the state of the buffers will be indicated by specific flags. The descriptor corresponding to the last occupied buffer will indicate the length of the frame.

The processing of the received Ethernet frames is achieved either through the callback mechanism or by repetidevely interogating the `ETH_Driver` class. The callback mechanism is similar to that of the SPI driver and supposes the implementation of the interface `ETH_Driver_Interface,` by the clients of the `ETH_Driver` class, and their registration through the method `b_AttachCallbackHandler(ETH_Driver_Interface* CallbackHandler).`

IV. Conclusions

The paper has presented a serial node based on the MPC5553 microcontroller. It is used for collecting data from sensors, through two SPI and two RS485 interfaces and sending them over the Internet.

Future development directions are:

- The extension of the performances of the existing interfaces, by increasing the communication speed and by allowing to connect multiple sensors on the same interface;
- Adding new interfaces; the focused interfaces are CAN, USB and for analogue inputs/outputs.

References

[1] H. Abrach, J. Carlson, H. Dai, J. Rose, A. Sheth, A, B. Shuker. and R. Han, “System Support for MultimodAINetworks of In-situ Sensors”, *Technical report CU – CS – 950 – 03,* April 2003, Available: http://iguanaworks.net/~shucker/papers/CU-CS-950-03.pdf

[2] J. Beutel, F. Mattern, K. Romer, F. Siegermund and L. Thiele, “Prototyping Wireless Sensor Network Applications with BT nodes”, *Lecture Notes in Computer Science,* Vol. 2920, 2004, Springer, pp. 323 – 338

[3] R. Dees and R. Marshall, MPC5500 family overview, *Freescale Semiconductor Engineering Bulletin,* No. 11, 2006, Available: http://www.freescale.com/files/32bit/doc/eng_bulletin/EB659.pdf

[4] A. Savvides and M. B. Srivastara, “Distributed Computation Platform for Wireless Embedded Sensing”, in *Proceedings of IEEE International Conference on Computer Design,* Freiburg, Germany, September 2002, pp. 220 – 225

[5] *** SPI description – Available : http://www.rpi.edu/dept/ecse/mpsd/SPI.pdf

[6] *** RS485 description – Available : http://www.rs485.com

EvalTool for Evaluate the Partitioning Scheme of Distributed Databases

Adrian Runceanu Marian Popescu
University Constantin Brâncuşi Târgu-Jiu, Romania
adrian_r@utgjiu.ro marian@utgjiu.ro

Abstract. **In this paper we present one tool named EvalTool for evaluate the partitioning scheme of distributed database. This tool could be use for calculate the best partition scheme both for horizontal and vertical partitioning, and then alocate the fragments in the right places over the network. We use an algorithm studied before, which could calculate some cost for access data to the local site and the other cost for access data from the remote site. The minimum value of this costs give the best fragmentation scheme for each relation.**

1. ABOUT FRAGMENTATION OF DISTRIBUTED DATABASES

The distributed databases design is an optimization process that requires to obtain solutions to several interpenetrating problems, namely data fragmentation, data allocation and local optimization. Each problem can be solved through different approaches, and therefore the design of distributed databases becomes a very difficult task. Usually the design process is heuristical defined.

The distributed databases design derived from the non-distributed databases design only in terms of distribution. The design involves data acquisition, database partitioning, the allocation and replication of partitions and local optimization. The database partitioning can be performed in several different ways: vertical, horizontal and hybrid (also called mixed) partitioning. This aspect of a database design will be highlighted in this thesis and namely the developing of an algorithm for checking some various algorithms proposed in literature for distributed databases partitioning (fragmentation).

Basically, it will be considered the problem of vertical data partitioning (fragmentation), also known as the attributes partitioning. This technique is used for the databases design to improve the performance of transactions. In vertical partitioning, the attributes of a relation R are together in groups that do not overlap and the relationship R is designed on relations fragments in accordance with these attributes groups.

In the distributed databases, these fragments are allocated on different sites. Here comes the aim of vertical partitioning to create vertical fragments of a relationship to minimize the cost of the data accessing during the transaction process. If the fragments are closer as possible to the needs of the transactions set, then the transactions processing cost can be reduced.

For the distributed databases design, the transaction cost is reduced by increasing the local transactions (on a site) and at the same time by reducing the amount of data views which are not local. The aim of vertical partitioning technique (and in general, partitioning techniques) is to find a partitioning scheme to meet the objective outlined above.

Note that the partitioning problem can be tackled on different levels of detail considering some additional information.

In paper [1], we have to consider only the information about transactions as input data to manage the partitioning problem effectively. In fact, the global optimization problem (which includes a large number of parameters and a metric complex) is divided into several smaller optimization problems in order to reduce the search space and to reduce the complexity of each problem separately.

In speciality literature are proposed several vertical partitioning algorithms, so it can be measured the affinity between pairs of attributes and try to group attributes together under the affinity between, using the algorithm BEA (bond energy algorithm). In an article we use an heuristical cost estimator for design a file to obtain a partitioning scheme "bottom-to-top". In another article extends the approach proposed by BEA and algorithm has two phases for vertical partitioning. The partitioning algorithms mentioned above use several heuristical methods for a relationship fragmentation.

As input date for these algorithms it is used the attributes matrix (AUM). This matrix has the attributes as columns, the transactions as lines and the transactions frequency access are values in the matrix. The most previous algorithms used for the data fragmentation use of affinity matrix of attributes (AAM) which is derived from AUM. AAM is a matrix in which for each pair of attributes it is stored the total frequency access of the transactions that access the pair of attributes. The results of different algorithms are often different, even if it is the same

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_109,

affinity matrix of the attributes, so it indicates that the objective functions used are different. Most vertical partitioning algorithms do not have an objective basis for assessing the correctness of partitions that are obtained through the application of those algorithms. Also, there is a common criterion or objective function to evaluate the results of these algorithms for vertical partitioning.

Bellatreche et al. [3] presents an horizontal fragmentation method and an analytical cost model to evaluate the query execution time in horizontally fragmented databases. The fragmentation schema with the best performance according to the cost model is achieved through a hill-climbing algorithm, by selecting a subset of classes for primary horizontal fragmentation.

The work by Ezeife and Barker [6] presents a set of algorithms for the horizontal fragmentation of all classes in a database schema. It takes relationships between classes into account to propose primary and derived horizontal fragmentation. However, this approach works at the instance level, where the class instances already exist in the database to proceed with the fragmentation process. It also assumes a storage structure for the objects hierarchy in the database class in which an instance of a subclass physically contains a pointer to the instance of its super-class that is logically part of it. This assumption leads to considering inheritance links in the horizontal fragmentation process.

Ceri et al. [4] proposes a methodology for the horizontal fragmentation of all classes in a database schema. The choice between primary and derived horizontal fragmentation on each class considers its relationships, which are defined by analyzing only the method calls between classes in the schema. The work does not present an algorithm to support the methodology.

Navathe et al. [8, 9] proposes a complete fragmentation methodology for OODBMS, which is divided in three phases. First, there is an Analysis Phase to assist distribution designers in defining the most adequate fragmentation algorithm (horizontal, vertical, or both) to be applied in each class of the database schema. The Analysis Phase also considers the case in which no fragmentation of a class is the best alternative. Second, they present an algorithm to perform Vertical Fragmentation in a class. Finally, the authors present an algorithm to perform Horizontal Fragmentation on the whole class or on a vertical fragment of a class, which may result in mixed fragmentation.

2. MAIN CARACTERISTICS

In few words EvalTool considers one relationship for time, once loaded is possible to fragment it (in the first time only vertically) and allocate the calculated fragments. It is also possible to load an external fragmentation and use it for the allocation, this might be useful as horizontal fragmentation.

For both the fragmentation and allocation, using the implemented algorithms, the tool reports all the possible solutions and the one considered the best.

The fragmentation and allocation algorithms are mainly based on the book [10]. We modified the cost analysis for the allocation, the model used is described in the next section.

The algorithms implemented are:

- Bound Energy Algorithm
- VF algorithm
- Best-Fit heuristic
- Cost Analysis (based on the model described in the next section)

The tool works with text files as input, at least it needs:

- one file for general information (sites names, applications frequencies, network values, and other information)
- one file for the information about a relationship (attributes names, useMatrix, and other information)

Once the 2 text files are loaded it is possible to vertical fragment the relationship. The tool works with the following steps:

- It is calculated the affinity matrix
- Clustering affinity matrix is calculated by the Bond Energy Algorithm (in the relationship text file is also possible to specify the column insertion order)
- all the solutions for splitting CA matrix into two fragments are evaluated with the VF algorithm
- if the best solution is to fragment the relationship, VF algorithm is iterated for each of the fragments calculated

The tool reports the CA matrix and all the steps of both the Bound Energy Algorithm and VF algorithm. Is also possible to store this information in a text file.

The allocation uses:

- the best fit heuristic
- a cost analysis (the model used is described in the next section)

EvalTool uses 2 policies of allocation:

- “one for each site” (the default one)
- “one for all sites” (can be activated selecting the relative option button)

An example will help to understand: consider that you are designing a distributed database for your university.

The university has one central site and five peripheral sites managing all the didactics.

Each peripheral site is independent from the others: just use the information in the database about its students, its exams, its teachers, and so one, it is easy that a few of the relationships of the database will be fragmented in one fragment for each peripheral site (e.g. the student relationship) and, because of the "independence" of the sites, the only solutions for allocation with sense will be to allocate the fragment on its site or on the central one: none of the other sites will use it; is easy that a lot of relationship will be fragmented in this way, think just to the students, the exams and the teachers relationships, each one fragmented in 5 fragments: not considering others fragmentation 15 fragments are created.

Using the "one for each site" (a more correct name would be "one fragment for each peripheral site") is possible to reduce the number of the fragments that the user has to manage: the input for the allocation won't be 3 text files with 5 fragments each one, but 3 text files with 1 fragment each one.

But there's also the possibility that a relationship is not fragmented in one fragment for each peripheral site (because is not possible to relate a tuple to a single peripheral site or simply because there's no need). In this case there aren't no-sense solutions for the allocation of the relationship (or the fragments obtained from other fragmentations), it can be allocating in all the sites central or peripherals; in this case must be used the "one for all sites" policy (one fragment for all sites). This policy can also be used if the user in the example before wants to create 3 text file with 5 fragments each one.

EvalTool was developed as part of a study for a grant research. Maybe the tool "fits too much" our problem so that there could be problems for using it in a more general one, but we considered useful to offer our solution to the community.

3. COST MODEL FOR ALLOCATION

We present an allocation model that minimizes the total cost of storage and processing, trying to comply with certain restrictions imposed on response time.

We consider a decision x_{kj} variable, defined as:

$$x_{kj} = \begin{cases} 1, & if \quad fragment F_k \quad is \quad placed\, on \quad site \quad S_j \\ 0, & in \quad the \quad other \quad cases \end{cases} \quad (1)$$

Total cost function has two components: processing applications and storage. This can be expressed as:

$$TOC = \sum_{\forall q_i \in Q} QPC_i + \sum_{\forall S_k \in S} \sum_{\forall F_j \in F} STC_{jk}, \quad (2)$$

where QPC_i is the cost of processing the application q_i and STC_{jk} fragment F_j is the cost of storage per station S_k. We can define the cost of storage with the formula:

$$STC_{jk} = USC_k * size(F_j) * x_{jk} \quad (3)$$

The cost of processing applications *(QPC)* can be separated into two components: the actual processing cost *(PC)* and the cost of transmission *(TQ)*.

$$QPC_i = PC_i + TC_i \quad (4)$$

Processing component consists of three cost factors: access cost *(CA),* the integrity constraint *(IC)* and control competitor *(CC)*:

$$PC_i = AC_i + IC_i + CC_i. \quad (5)$$

Detailed specification of these three cost factors depend on the algorithms used for these tasks. For example, *AC* has the form:

$$\sum_{\forall F_j \in F} STC_{jk} \leq storage capacity \; at \quad site \quad S_k, \quad \forall S \in S. \quad (6)$$

The first two terms from the previous formula calculates the number of accesses to the query qi the fragment Fj and UR + RR value provides the total number of accesses that executes work | retrieve and update them. Local cost of processing them is assumed to be identical, given LPCk multiplying the total cost of access to station Sk and multiplying xjk show that will select only those values of the cost for stations where parts are stored.

Access cost function requires that each application can be divided into subqueries that run on each fragment stored on a station, the results are returned to the station from which the request was sent. This is a simplistic view that ignores the complexity of database processing. For example, it was revealed the cost of execution join unlor or cost competitiveness and integrity.

The cost of transmission can be made analogous to the cost of access. The queries that are updating is required to inform all stations where the replicas, while the queries that are restored, it is sufficient to access only one of the descendants, in addition, after the application update, no data transmission back to state of origin (except a confirmation message), while the queries that are restored may result in transmission of data. Component update transmission function has the form:

$$TCU_i = \sum_{\forall S_k \in S} \sum_{\forall F_j \in F} u_{ij} * x_{jk} * g_{o(i),k} + \sum_{\forall S_k \in S} \sum_{\forall F_j \in F} u_{ij} * x_{jk} * g_{k,o(i)}, \quad (7)$$

where the first term is to send a message to update the initial station 0 (0 demand all copies of fragments that must be updated and the second term is to confirm transmission). Component retrieve can be specified as follows:

$$TCR_i = \sum_{\forall F_j \in F} \min_{S_k \in S} (r_{ij} * x_{jk} * g_{o(i),k} + r_{ij} * x_{jk} * \frac{sel_i(F_j)}{fsize} * g_{k,o(i)}). \quad (8)$$

The first term in TCR is the cost of forwarding the application to recover all stations contain copies of the fragments to be accessed, and the second term is the cost of transmission results from these stations to the original station.

The cost of transmission for q_t query can be described as:

$$TC_i = TCU_i + TCR_i \qquad (9)$$

Constraint functions can be analyzed similarly. For example, response time constraint can be described thus: *execution time of the call* $q_i \leq$ *maximum response time of applications qi,* $\forall q_i \in Q$, and storage constraints may be described by the relation:

$$AC_i = \sum_{\forall S_k \in S} \sum_{\forall F_j \in F} (u_{ij} * UR_{ij} + r_{ij} * RR_{ij}) * x_{jk} * LPC_k, \quad (10)$$

The heuristic methods must be sought that produce sub-optimal solutions and must test the closeness of the results of heuristic algorithm for optimal allocation. Unfortunately there is insufficient information to determine the “near” optimal solution.

4. PARTITION EVALUATION FUNCTION

The partition evaluation function is an objective function used to evaluate the "goodness" of a particular partitioning scheme. It can be used to measure the performance of a fragmentation system whenever a major change in access pattern information or global conceptual schema occurs. However, the defining of the suitable function for the vertical fragments and the building of EvalTool for vertical fragmentation are the main contributions of this work. [1]

5. THE PARTITIONING EVALUATION FUNCTION

Attributes of relations are the members of relational vertical fragments. On the other hand, the methods and attributes of object classes are the member of object vertical fragments. The instances are members of the horizontal fragments object. The set of instance objects in fragments created by different fragmentation schemes may be different. It is necessary to develop a modified partition evaluator to deal with object horizontal fragmentation schemes. This partition evaluator counts the number of local irrelevant accesses made to fragments by queries as well as the numbers of remote accesses made to fragments for each set of horizontal fragments created with a scheme.

Therefore, in order to measure the system performance each fragment is assumed placed at one of the distributed sites, while the PE function is used to measure the partition evaluation value (the PE value) for the system. The higher the PE value of the system, the worse the system performance. Computing the PE value of horizontal fragments requires an input matrix that shows for each application the number of times an instance object is accessed. This is called the application instance matrix. In [5], Chakaravarthy et al. computes the processing costs of distributed fragments using an application/attribute matrix that counts the number of times applications access attributes at distributed sites.

This paper computes the partition evaluator value (PE) with this formula:

$$PE = E_L^2 + E_R^2 \qquad (11)$$

where

$$E_L^2 = \sum_{i=1}^{M} \sum_{t=1}^{T} \left[q_t^2 * |S_{it}| \left(1 - \frac{|S_{it}|}{n_i} \right) \right] \quad (12)$$

n - Total number of attributes in a relation that is being partitioned.
T - Total number of transactions that are under consideration.
q_t - frequency of transaction t for t = 1,2, …, T.
M - Total number of fragments of a partition.
n_i - number of attributes in fragment i.
S_{it} - set of relevant attributes in fragment i that the transaction t
accesses; it is empty if t does not need fragment i.
$|S_{it}|$ - number of relevant attributes in fragment i that the transaction t accesses.

$$E_R^2 = \sum_{t=1}^{T} \min_{i=1}^{M} \sum_{k \neq i} \left[\sum_{P_k} \left(f_{tp_k}^{k} \right)^2 * \frac{|R_{itk}|}{n_{itk}^{remote}} \right] \quad (13)$$

n_i – number of attributes in fragment i.
n_{itk}^{remote} - Total number of attributes that are in fragment k accessed remotely with respect to fragment i by transaction t.
f_{tj}^{i} - Frequency of transaction t accessing attribute j in fragment i
R_{itk} - Set of relevant attributes in fragment k accessed remotely with respect to fragment i by transaction t; these are attributes not in fragment i but needed by t
$|R_{itk}|$ - Number of relevant attributes in fragment k accessed remotely with respect to fragment i by transaction t

The algorithm implements the objective function defined above. For measuring the performance of a partition scheme, algorithm needs two inputs. One is every horizontal fragment from the partition scheme; the other is the application object instance sets. Then produces the PE value of this partition scheme. Each PE value corresponds to the total penalty cost incurred by each scheme through both local irrelevant access costs and remote relevant access costs. A lower PE value means less performance penalty and thus, a better performance. The algorithm is below. This algorithm accepts the application object instance sets of al1 applications, the set of horizontal fragments allocated to distributed sites and application access frequencies at distributed sites. It then computes the local irrelevant access cost of each class as the sum of local irrelevant access cost of al1 applications for each fragment of the class. Similarly, it computes the remote relevant access cost as the sum of remote access cost made by al1 applications to fragments of this class at remote sites. The sum of these two costs makes up the PE value of the class.

We implemented the algorithm based on the formula from equation (11) calculates the value of PE, for a given fragmentation scheme so we used an entrance date: the matrix used for attributes - A; the lots of fragments on which it calculated the value of PE, the relation -R.

Algorithm

Input: A = attribute usage matrix;
R = Relation ; F = fragments set

Output: E^2_L: irrelevant local attribute cost;
E^2_R: relevant remote attribute cost;
EP : partition evaluator value

Begin
$E^2_L=0$
for i **from** 1 **to** M **do**
$e_i=0$
begin
for j **from** 1 **to** T **do**
$e_i=e_i+q_t^2*|S_{it}| * (1 - |S_{it}| / n_i)$
end_for
$E^2_M = E^2_M + e_i$
end_for
$E^2_R=0$
for t **from** 1 **to** t **do**
begin
minim =maxint
for i **from** 1 **to** M **do**
for k **from** 1 **to** M **do**
begin
if $k \neq i$ **then**
begin
if exist attribute in matrix A who is from k fragment **then**
$E^2_R= E^2_R+ (f^t_k)^2 * | R_{it} | * | R_{it} | / n^{remote}_{itk})$
end_if
if $E^2_{R\ min}$ < minim **then**
minim = $E^2_{R\ min}$
end_if
end_for
$E^2_R = E^2_R + E^2_{R\ min}$
end_for
end_for
$PE = E^2_L + E^2_R$
End.

6. EXPERIMENTAL RESULTS

For the execution of one transaction, we know that if a transaction could be run at one fragment and that fragment haven't one single attribute accessed by that transaction, then transaction not be run on that fragment. We consider 3 cases of matrix of attributes.

For the first test we used a matrix of attributes use with ten attributes accesed by eight transactions.

Tranzactions \ Attributes	1	2	3	4	5	6	7	8	9	10
*T*1	25	0	0	0	25	0	25	0	0	0
*T*2	0	50	50	0	0	0	0	50	50	0
*T*3	0	0	0	25	0	25	0	0	0	25
*T*4	0	35	0	0	0	0	35	35	0	0
*T*5	25	25	25	0	25	0	25	25	25	0
*T*6	25	0	0	0	25	0	0	0	0	0
*T*7	0	0	25	0	0	0	0	0	25	0
*T*8	0	0	15	15	0	15	0	0	15	15

We present in Figure 1 values for each number of fragments and values for E^2_M, E^2_R and EP.

Total number of fragments evaluates was 115975. Optimal value (minimum) is obtained for 3 fragments – fragment I (1,5,7), fragment II (2,3,8,9) and fragment III (4,6,10).

The program to generate all the combinations of ten attributes accessed by eight transactions offers three solutions (for five fragments) and two solutions (for eight fragments), having the same value for EP. However, the project of distributed database can choose which scheme of partition wishes to use it.

Number of fragments	**Partition scheme**	E^2_M **values**	E^2_R **values**	EP **values**
1	(1,2,3,4,5,6)	24895	0	24895
2	(1,4) (2,3,5,6)	7565	55	7620
3	(1) (2,3,5,6) (4)	7565	276	7841
4	(1) (2) (3,5,6) (4)	5063	11336	16399
5	(1) (2) (3,6) (5) (4)	0	22492	22492
6	(1) (2) (3) (4) (5) (6)	0	40913	40913

Fig. 1 Results for test

7. CONCLUSIONS

In this paper is presented a general approach of the fragmentation issue of the dates from a distributed database. Using this tool for obtained the best partitioning scheme with implementation of several clasic algorithms is one solution in designing phase of a distributed database.

REFERENCES

[1] Runceanu A. *Fragmentation in distributed databases*, (2008), International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 2007) Conference, Conference Proceedings book, December 3-12, 2007, University of

Brigeport, USA, publish in *Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, Springer Science*, ISBN 978-1-4020-8734-9 (Print) 978-1-4020-8735-6 (Online), DOI 10.1007/978-1-4020-8735-6_12.

[2] J. Banerjee, W. Kim, and K.C. Kim, "Queries in object oriented databases," in Proceedings of the IEEE Data Engineering Conference (ICDE'88), February 1988, pp. 31–38.

[3] L. Bellatreche, K. Karlapalem, and A. Simonet, "Horizontal class partitioning in object-oriented databases," in proceedings of the 8th International Conference on Database and Expert Systems Applications (DEXA'97), September 1997, pp. 58–67, Lecture Notes in Computer Science 1308.

[4] S. Ceri, M. Negri, and G. Pelagatti, "Horizontal data partitioning in database design," in Proceedings of the ACM SIGMOD International Conference on Management of Data, 1982, pp. 128–136.

[5] Chakravarthy S., Muthuraj R., Varadarajan R., and Navathe S. (1994) An objective function for vertically partitioning relations in distributed databases and its analysis. In *Distributed and parallel databases*, pages 183-207, Kluwer Academic Publishers.

[6] C.I. Ezeife and K. Barker, "A comprehensive approach to horizontal class fragmentation in distributed object based system" International Journal of Distributed and Parallel Databases, vol. 3, no. 3, pp. 247–272, 1995.

[7] W. Kim, *Introduction to Object-Oriented Databses*, MIT Press, 1990.

[8] S.B. Navathe, K. Karlapalem, and M. Ra, "Amixed partitioning methodology for distributed database design," Journal of Computer and Software Engineering, vol. 3, no. 4, pp. 395–426, 1995.

[9] S.B. Navathe and M. Ra, "Vertical partitioning for database design: A graphical algorithm," in Proceeding of the International Conference on Management of Data, ACM-SIGMOD'89, 1989, pp. 440–450.

[10] Tamer O. and Valduriez P. (1999) *Principles of Distributed Database Systems*. Prentice Hall Englewood Cliffs, Second Edition, New Jersey 07362.

[11] Grofu F., Popescu L. and Cercel C. (2008) Acquisition System for Monitoring Vibrations, International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 2007) Conference**,** Conference Proceedings book, December 3-12, 2007, University of Brigeport, USA, publish in Novel Algorithms and Techniques In Telecommunications, Automation and Industrial Electronics, ISBN 978-1-4020-8736-3 (Print) 978-1-4020-8737-0 (Online), pg. 89-92, DOI 10.1007/978-1-4020-8737-0_16.

Automatic Creation of Daily News Digests from Internet News Sites

David B. Bracewell
General Electric Global Research
Niskayuna, NY 12309
bracewed@ge.com

***Abstract*—With the advent of 24-hour news channels and Internet news sites, the ease of finding news sources has become easier than ever. However, with the increased number of sources covering a greater variety of topics it is becoming increasingly difficult for users to weed through the important and not so important news of the day. This paper presents a system that automatically creates a daily digest of the important news topics of the day. News articles are mined from Internet news sites and the articles are stored for later analysis. At the end of the day, or specified period, the articles are analyzed to determine the important news topics and to create a listing of the top stories of the day. The digest shows the top stories with their title, keywords, link and summary information. Through human judged evaluation it is shown that the created digests satisfy users.**

Keywords-component; Digest Creation, Online News, Personalization, Synopsis Creation

I. Introduction

In the past 10 years the number of online news sources has grown tremendously. With the increase in sources have also come increases in the number of news articles, topics covered, etc. that are available. Keeping adequately informed, in an era of information overload, requires constant attention to Internet news sites, the sites' RSS feeds or backtracking at the end of the day to determine the important events.

Online news has become a hot topic in the information retrieval world. From workshops on topic detection and tracking to various research projects, the information retrieval of online news is of growing importance. There has been work done to automatically create RSS feeds of news archives [4] to make past news more easily available. There have also been efforts to combine news, such as the Pan-European news distribution network [8]. However, keeping track of important and popular stories and topics is still a problem.

In order to help the user find popular stories that may be of interest or may be important, some online news sites offer most recommended and most viewed sections. This allows for users to quickly scan articles that others have found interesting. In addition, news sites are making good use of RSS feeds that allow users to view only specific content. Another way of helping users is to offer daily digests. A digest summarizes the most important topics and stories of the day. Some sites are offering daily digests, which are typically manually created. However, there seems to be little research done on automatically creating digests.

This paper presents techniques for gathering and classifying news articles and for creating simple digests. The system is based on a keyword extraction algorithm that is applicable to multiple languages. Currently, we are focused on creating an English and Japanese version of the system.

This paper will continue as follows. In section 2 a brief overview of the proposed system will be given. Then, in section 3 the news gathering and classification module will be discussed. In section 4 an overview of creating daily digests from the classified news articles is given. Then, in section 5 a look at the current implementation details will be given. Section 6 shows evaluation results. Finally, in section 7 concluding remarks will be made and future work discussed.

II. System Overview

The system is made up of two modules: the news monitoring and article processing module and the digest creation module. The news monitoring and article processing module monitors a list of user defined RSS feeds and news site pages for new news and the downloads, extracts and processes the articles. Processing the articles includes keyword extraction, category classification and summary generation. The digest creation module takes the downloaded and processed articles at the end of the day and creates a digest of the news by determining which topics and articles were the most important for the day. An overview can be seen in figure 1.

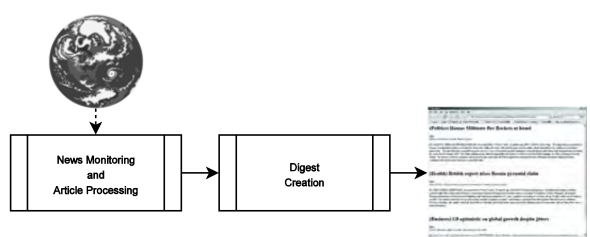


Figure 1. System Overview

The system is designed to be automatic and does not require any intervention by end users. The only requirement from the end user is for them to set the time at which the news monitoring and article processing module runs. For example,

This research was originally done while the author was at The University of Tokushima.

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2_110,

the user could set the module to run from 6am to 5pm everyday. At completion, the digest creation module would take over and create a digest for the time span given by the user.

The digest creation module outputs in HTML, which could be easily used as a digest, web page served on the Internet. The main output of the system is the digest, which is made up of the top articles of the day with their category or genre (sports, business, politics, etc.), top keywords and a summary. Additional output includes ranked lists of keywords and categories for the day.

III. News Monitoring and Article Processing

News monitoring and article processing handles monitoring, downloading and processing news articles during an assigned time span. An overview of the module can be seen in figure 2. The module monitors a list of RSS feeds and news web sites for new news. After articles are downloaded and extracted they are processed by extracting keywords, assigning categories and generating summaries.

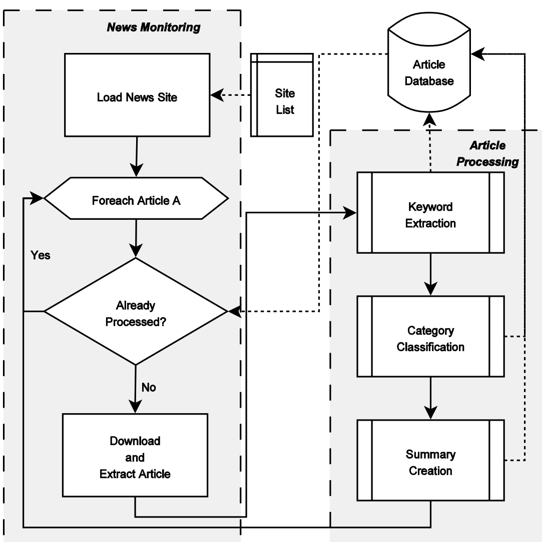


Figure 2. Gathering and Classification Overview

A. *News Monitoring*

The news monitoring part of the module watches a set of RSS feeds and news site web pages for new news articles. The sites are processed one at a time in sequential order. Depending on the site either an RSS feed or HTML page is parsed and the links and article titles are extracted.

After the articles are collected they are processed one at a time. First, the article's link and title are checked for in the article database, which stores the article and its processing results. If the link or the title is in the database then the article is treated as already processed and is ignored. If the article has not been processed then the link is used to download the page containing the article. The article is then extracted from the HTML web page using a rule-based extraction process.

B. *Article Extraction*

After examining many news sites we found a certain pattern that described the prototypical article. The pattern is made up of a title, then the article body and the footer. Certain sites, however, utilize handcrafted rules in the form of regular expressions that allow for very precise extraction. However, this poses a possible bottleneck for the addition of new news sites.

The first step in article extraction is to find the title in the body of the HTML. This is done using the title extracted from the RSS feed or web page. Since these titles sometimes have extraneous information, such as web site name or provider, they cannot be fully matched to a line in the HTML. Instead, each line is examined to see if it is a substring of the title. If it is, then that line is marked as the article's title.

The next step is to find the article footer. This is done by examining each line that comes after the article's title to see if it contains one of the defined footer elements. Some of the currently used footer patterns are Copyright, ©, (C), Email this, Related Stories, etc. If no footer pattern is found then the bottom of the HTML body is used.

After the footer is found the article can be extracted. The article body is the text that falls between the title and the footer. The article is then cleansed. The cleansing process removes multiple white space and unimportant multiple punctuation marks, such as --, **, ..., etc.

C. *Article Processing*

Article processing consists of keyword extraction, category classification and summarization. Keyword extraction extracts the most important words and phrases for describing the article and its content. Category classification assigns to each article one or more predefined news categories. Summarization gives the important points of the article in a few sentences.

1) *Keyword Extraction*

For keyword extraction, the method proposed by Bracewell et al. [1] was used. It is capable of extracting keywords from a single document without the need for training data or corpus statistics. A brief description will be given in the next couple of paragraphs.

The algorithm uses linguistic information in the form of noun phrases to extract keywords. Phrases can make better keywords than single words as they can keep concepts portrayed in compound words together. For example ``New York'' would be a better keyword than having ``New'' and ``York'' separately.

The algorithm is made up of three modules: morphological analysis, NP extraction and scoring, and NP clustering and scoring. Morphological analysis takes care of word segmentation, part-of-speech tagging, and word stemming. For English, Porter's stemmer [7] and Brill's tagger [3] were used. For Japanese, Chasen [6] was used.

NP extraction and scoring extracts simple noun phrases and gives them a score based on their frequency in the article and the frequency of the words that make it up in the article, see [1] for details on the score. NP clustering and scoring brings together noun phrases that share some commonality and is an attempt to prevent redundancy in the final set of keywords. For example, "bank bailout" and "bailout" would be clustered together. The clusters are scored based on the scores of the noun phrases in them. From each of the clusters a representative is chosen and becomes a keyword for the article.

2) Category Classification

For category classification, we used the method proposed in [2], which is based on the previously mentioned keyword extraction algorithm. It is a keyword based classification method that only requires positive examples. It has the added ability that it is capable of updating the training information in an online manner as well as easily adding new categories without a complete retraining. It also had an F-Measure of 93% on the test set in that paper. A brief description will be given in the next couple of paragraphs.

The algorithm is made up of N independent classifiers, where N is the number of categories. Training involves building a model from a set of articles that are in the desired category. A category model is made up of a category name, total number of documents counter and a keyword list. The keywords list contains the keywords and their raw document frequency for the category.

The classifiers output the likelihood that an article is in the category. The likelihoods from all the categories are used to determine which of the categories should be assigned to the article. The categories that have a likelihood greater than the mean plus one standard deviation are assigned to the article. The assumption is that these categories stand out among the group and are the best choices for the article.

3) Summarization

For summarizing the article, the extracted keywords are used to assign a score to each of the sentences. The top ***N*** sentences are then extracted and become the article's summary. In this research ***N*** = 4. The idea behind the summarization technique is that the sentences with a high number of keywords in them are more important than sentences that have a lower number or no keywords.

Summarization is done in three steps. The first step is to break the article up into individual sentences. This is done using a rule-based approach. The second step is to score each of the sentences using equation 1, where $k_1, \ldots, k_n$ are the keywords in sentence ***S*** and ***KScore(k_i)*** is the keyword score of k_i. The final step is to create the summary. For doing this, the sentences are first sorted by score and the top four extracted. Then the four sentences are sorted by position in the article and used as the summary.

$$Score(k_1 \ldots k_n \in S) = \sum_{i=1}^{N} \log(1 + KScore(k_i)) \qquad (1)$$

IV. DIGEST CREATION

After the current day's news has been downloaded and processed, a digest can be created that will highlight the important news of the day. The digest creation module is made up of two parts: article set determination and synopsis creation. Article set determination uses clustering in an attempt to determine the broad topics covered in the day's articles. Synopsis creation takes the top topics and determines the representative article from them.

A. Article Set Determination

The first step in creating the digest is to determine which set of articles were the most important of the day. This is done by clustering the articles and ranking the resulting clusters. The overview of this can be seen in figure 3.

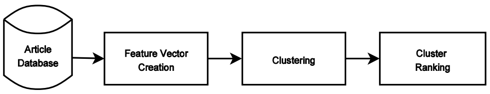


Figure 3. Article Set Determination Overview

First, a feature vector is created for each article. The feature vector is simply a vector space model representation of the article using the keywords that were extracted in the new monitoring and article processing module. The next step clusters the articles in an attempt to group articles of similar topic together.

K-Means clustering [5] is used and the distance between two articles is calculated as one minus the cosine similarity, shown in equation 2. The value of K is automatically estimated in the following manner. First, the range of possible K is determined by using the lower bound of ***N/2*** where ***N*** is the number of articles downloaded and an upper bound ***N***.

$$\text{Distance}(A_i, A_j) = 1 - \frac{A_i \bullet A_j}{|A_i| \times |A_j|} \qquad (2)$$

Five random, but unique, values are chosen from the range. Each of these values are chosen to represent K and the K-Means algorithm is allowed to run for a few steps. After completion, the average intra-cluster and extra-cluster distance are calculated. The value with the best combination of low intra-cluster and high extra-cluster distance is chosen and the K-Means algorithm is run until completion or until 100 steps have been completed.

$$CScore(Cluster) = P(Category) \times |Cluster| \qquad (3)$$

After the articles are clustered and have a category assigned, they are ranked. The category chosen for the cluster is the most abundant one among the articles in the cluster. For ranking each cluster is first scored using equation 3. The probability of a category, ***P(Category)***, is estimated using the frequencies of the classified categories for the day's articles and as such, will vary day to day.

B. *Synopsis Creation*

After the clusters are ranked by cluster score, the top clusters can be chosen as the most important ones. Currently, the top 10 clusters are used, but the user could easily change this. For each of these clusters an article is chosen to represent the cluster. To choose the representative, article the importance of each article is calculated using equation 4, where $\boldsymbol{A}$ is the set of keywords that define the article, $\boldsymbol{DF(k_i)}$ is the number of articles the keyword appeared in, and $\boldsymbol{KScore(k_i)}$ is the keyword score of keyword $\boldsymbol{k_i}$.

$$\text{Importance}(A) = \sum_{i=1}^{N} DF(k_i) \times KScore(k_i) \qquad (4)$$

For the top clusters, a brief synopsis is created. The synopsis consists of the cluster's category, the representative article's title, summary and link to the full article, and the top 5 highest scoring keywords in the cluster.

V. IMPLEMENTATION

The system has been developed in C# and MySQL is used for storing the articles and their classification results. Currently, English and Japanese news is being monitored and gathered using a set of 12 RSS feeds from the news sites listed below.

- Yahoo! Top Stories (English)
- Yahoo! Most Viewed (English)
- Google News (English)
- Washington Post Top News (English)
- Nikkeibp.jp News (Japanese)
- Asahi (Japanese)
- Livedoor Main Topics (Japanese)

Eight categories, business, crime and misfortune, politics, sports, health, entertainment, science and nature, and technology, are being used for category classification. Each category was trained using a collection of 1,000 documents representing the topic. Each category has a Japanese and an English version.

The news monitoring and article processing module processes the feeds every hour. This is, currently, resulting in acquiring and processing about 1,200 articles a day. Because of the large amount of articles that are being collected, the news digest creation algorithm is taking upwards of an hour to complete. Figures 4-6 show examples of output from the system.

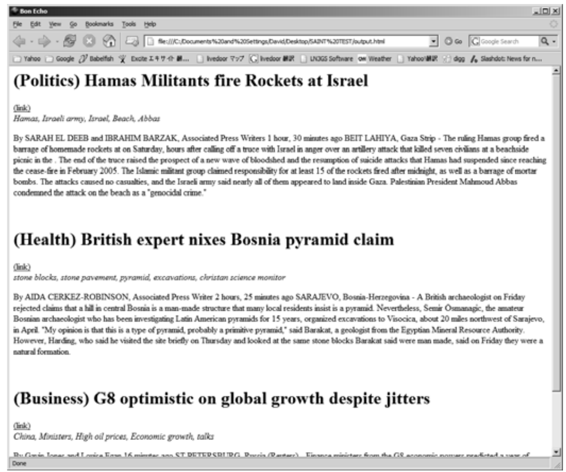


Figure 4. Digest Example

Politics	0.43
Business	0.21
Science & Nature	0.15
Crime & Misfortune	0.14
Entertainment	0.13
Sports	0.12
Health	0.09

Figure 5. Category Ranking Example

minut	0.20
recent	0.18
violenc	0.18
tool	0.18
video	0.17
warner	0.17
read	0.17
lllp	0.17
reserv	0.17
cabl network	0.17

Figure 6. Keyword Ranking Example

VI. EVALUATION

The proposed system was evaluated by comparing the digests to human chosen important topics. The topics were taken from titles of articles appearing during the day and manually cleaned up to be more generic. Examples of topics are “Colts win Superbowl” and “Abe elected prime minister.” Forty topics were created for each day over a one-week period starting March 13th, 2007 taken from the 10 stories chosen by the proposed system and 30 other frequent topics found. In addition to these topics, judges were told to write-in their own topics if they did not appear in the list.

Evaluation was done for English and Japanese looking at the agreement for the top 10 stories of the day. Each language had two judges.

English results can be seen in Table 1. Judge 1 and Judge 2 had very different agreements. Judge 2 leaned more in favor of political articles, while judge 1 more world events. The use of personalization techniques could be used to improve the individual results.

TABLE I. ENGLISH EVALUATION RESULTS

Judge	Agreement
1	26%
2	78%
$1 \cap 2$	83%

Japanese results can be seen in Table 2. Judge 1 and Judge 2 had a greater similarity in agreement than judges for English. Both judges leaned towards political and business articles.

TABLE II. JAPANESE EVALUATION RESULTS

Judge	Agreement
1	43%
2	41%
$1 \cap 2$	49%

VII. CONCLUSION

In this paper a system that creates daily digests of news articles was presented. The system monitors and gathers news to download and classify. Classification is made up of keyword extraction, category classification and summarization. Using all of this information a digest can be created that highlights the important news of the day as well as the top categories and keywords. Currently, the system works on English and Japanese news.

In the future, we hope to configure the system to work on the web and add more news sources. We would also like to add news tracking, which would track stories over time. It would also be interesting to add named entity recognition to find the top people and places in the news for the day. We would like to also create a Chinese version of the system. Finally, we want to explore ways to improve the system. Personalizing the digests per user based on user defined topics of interest and feedback could improve the agreement. We hope to examine a large number of human created digests to mine features of digest-worthy articles.

REFERENCES

[1] D. B. Bracewell, F. Ren, and S. Kuroiwa. Multilingual single document keyword extraction for information retrieval. In Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Wuhan, China, November 2005.

[2] D. B. Bracewell, F. Ren, and S. Kuroiwa. Category classification and topic discovery of news articles. In Proceedings of Information-MFCSIT 2006, Cork, Ireland, August 2006.

[3] E. Brill. A simple rule-based part-of-speech tagger. In Proceedings of 3rd Applied Natural Language Processing, pages 152–155, 1992.

[4] S. Debnath. Creating rss for news articles and beyond. In Proceedings of the 19th FLAIRS Conference, pages 275–280, 2006.

[5] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297, 1967.

[6] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. Morphological analysis system chasen version 2.2.9 manual. Technical report, Nara Institute of Science and Technology, 2002.

[7] M. Porter. An algorithm for suffix stripping. Program, 14:130–137, 1980.

[8] M. Schranz, S. Dustdar, and C. Platzer. Building an integrated pan-european news distribution network. In Proceedings of the 6th IFIP Working Conference on Virtual Enterprises, PRO-VE 2005, pages 26–28, 2005.

Index

K. Elleithy et al. (eds.), *Technological Developments in Networking, Education and Automation*,
DOI 10.1007/978-90-481-9151-2,

P

Q

R

S