Decision-theoretic Image Retrieval with Embedded Multi-resolution Mixtures

Nuno Vasconcelos

Cambridge Research Laboratory

Technical Report Series

CRL 2002/04

April 2002

COMPAQ

Decision-theoretic Image Retrieval with Embedded Multi-resolution Mixtures

Nuno Vasconcelos
Cambridge Research Laboratory
Compaq Computer Corporation
Cambridge MA 02139

April 2002

Abstract

The design of an effective architecture for image retrieval requires careful consideration of the interplay between the three major components of a retrieval system: feature transformation, feature representation, and similarity function. We introduce a decision theoretic formulation of the retrieval problem that enables the design of systems where all components are optimized with respect to the same end-to-end performance criteria: the minimization of the probability of retrieval error. The new formulation is shown to have two appealing properties. First, it leads to an optimal similarity function (the posterior probability of the query under the database image class) that generalizes many of its previously proposed counterparts. Second, it enables a theoretical characterization of the impact of the feature transformation and representation in the probability of error. In addition to exposing the major limitations of a large body of previous retrieval approaches, this characterization allows the derivation of a series of conditions for the optimal design of the feature transformation and representation. The search for a practical solution that can satisfy these conditions leads to the adoption of an embedded multi-resolution mixture representation and originates an efficient algorithm for optimal feature selection. The resulting retrieval architecture achieves a good compromise between retrieval accuracy, invariance, perceptual relevance of similarity judgments, and complexity. Extensive experimental results show that decision-theoretic retrieval performs well on color, texture, and generic image databases in terms of both retrieval accuracy and perceptual relevance of similarity judgments.

Author email: nuno.vasconcelos@compaq.com

CRL Technical reports are available on the CRL's web page at
http://crl.research.compaq.com.

Compaq Computer Corporation
Cambridge Research Laboratory
One Cambridge Center
Cambridge, Massachusetts 02142 USA

1 Introduction

An architecture for content-based image retrieval (CBIR) consists of three fundamental building blocks: 1) a feature transformation from the space of image observations (e.g. pixels) to a feature space with better retrieval properties, 2) a feature representation that compactly describes how each of the database image classes populates this space, and 3) a similarity function that allows ranking the database classes by similarity to a query. While significant attention has been devoted to each of these individual components, there have been significantly fewer attempts to investigate the interrelationships among them and how these relationships affect the overall performance of retrieval systems.

In fact, a significant fraction of the retrieval literature can be classified into two major groups, according to the emphasis placed on the design of the individual retrieval components. The first group contains solutions tailored for texture (to which we refer as *texture-based retrieval*) while the second contains solutions tailored for color (*color-based retrieval*). Texture retrieval approaches tend to place all emphasis on the design of the feature transformation. The key idea, which can be explicit in the formulation of the problem [71, 75, 17] or only implicit [40, 45, 44, 46], is to find *discriminant* feature transformations. These are transformations that best separate the feature distributions of the different image classes. Ideally, given small class overlap, simple similarity metrics such as the Euclidean or the *Mahalanobis distance* (MD) should guarantee good retrieval performance.

On the other hand, discrimination has not been a critical issue for color-based retrieval, where the features are either the pixel colors themselves or color-ratios that guarantee different types of invariance [23, 26]. Instead significant work has been devoted to the feature representation, consisting mostly of variations on the color histogram [68], e.g. the color coherence vector [52], the color correlogram [28], color moments [66], etc. Here, similarity metrics are usually L^p norms and, among these, the L^1 distance, also known as *histogram intersection* (HI) [68], has become quite popular [68, 56, 58, 41, 59, 64, 66, 1].

While they have worked reasonably well in their specific domains, these representations break down when applied to generic databases. On one hand, the discriminant transformations proposed by texture-based approaches tend to be database specific, e.g. discriminant features for a texture database are usually not discriminant for an object database, and it is therefore not clear that such approaches can be generalized to the full-blown retrieval problem (where image content is unconstrained). On the other, color-based solutions are plagued by the exponential complexity of the histogram on the dimension of the feature space, and are therefore only applicable to low-dimensional feature spaces (e.g. the space of pixel colors). Hence, they are unable to capture the spatial dependencies that are crucial for characterizing image properties such as texture or local surface appearance.

The alternative to concentrating on the features or feature representation, is to investigate the design of retrieval systems that are optimal in some end-to-end sense, i.e. where all retrieval components are optimized with respect to the same overall performance criteria. This, of course, raises the problem of defining a meaningful criteria for end-to-end optimality. Since the ultimate goal of any retrieval system is to be correct as often as possible, we formalize the retrieval problem as one of decision theory

and adopt the criteria of *minimizing probability of retrieval error* (MPE). The decision-theoretic formulation has two major properties of interest. First, it leads to *generic* solutions which are optimal in a sense (MPE) that is meaningful for any type of visual databases, e.g. object databases, texture databases, databases of consumer photographs, and so forth. Second, it makes a vast body of existing decision-theoretic results relevant to the retrieval problem, simplifying the task of designing optimal systems.

One well known such result is that the optimal similarity function, in the MPE sense, is that associated with the *Bayes classifier*: the posterior probability, under each database class, of the features vectors in the query. In this work, Bayesian similarity is 1) shown to generalize many of the similarity functions (Mahalanobis distance, χ^2 statistic, and minimum discrimination information, among others) in common use in the retrieval literature, and 2) used as a starting point for a decision-theoretic analysis of the trade-offs to be satisfied by the retrieval components, when the goal is to achieve end-to-end optimality. The main result of this analysis is that any retrieval system must indeed achieve a compromise between feature transformation and feature representation, taking into account three conflicting constraints:

- fine image discrimination requires the ability to capture local dependencies between image pixels, which can only be achieved through spatially supported features, i.e. when the space of image observations is high-dimensional.
- (MPE) optimal performance is only guaranteed for a restricted set of *invertible feature transformations*;
- density estimates tend to be poor when the feature space is high-dimensional.

Because an invertible transformation can only map a high dimensional observation space into a high-dimensional feature space, where it is difficult to obtain reliable density estimates, it follows that the design of decision-theoretic retrieval systems always requires either sacrificing the invertability of the transformation (allowing the feature space to be low-dimensional even when the observation space is not), or sacrificing the spatial support of the features (by relying on low-dimensional observation spaces).

Since either of these can have drastic consequences on retrieval accuracy, it is important to base design decisions on a solid understanding of all the involved trade-offs. To obtain such understanding we introduce the notion of *embedded feature spaces*, which are the spaces obtained by sequential downward projection of a starting feature space. Embedded feature spaces are shown to be an *intrinsic* component of retrieval systems with linear feature transformations, in the sense that any such transformation originates a sequence of embedded spaces with monotonically decreasing lower bound in probability of error and monotonically increasing density estimation error. In result, for a given feature transformation, the probability of error is a convex function of the number of embedded subspaces considered in the retrieval operation. It follows that the problem of optimal feature design can be decoupled into two smaller subproblems: 1) finding the best invertible feature transformation, and 2) finding the subspace dimension where the probability of error achieves its minimum value.

In general, these are difficult problems which involve iterating between density estimation and feature updating, two steps that must cycle through all image classes in

the database. We show, however, that efficient solutions are possible whenever the set of transformations of interest is finite, the search restricted to sequences of embedded subspaces of a common transformation, and the Gauss mixture adopted for feature representation. The latter is a particularly interesting result due to the fact that Gauss mixtures exhibit three other properties that are appealing in the retrieval context: computational tractability in high-dimensional spaces, ability to approximate arbitrary densities, and compactness. Overall, this leads to the notion of a sequence of *embedded mixture models*. Given a mixture density defined on a starting feature space, this is simply the sequence of mixtures resulting from the projection into the associated embedded subspaces.

Once the Bayesian similarity criteria and the embedded mixture representation are in place, it remains to determine the best finite set of feature transformations to consider during feature design. Here we simply draw on what is known about the human visual system and consider the set of multi-resolution transforms. This leads to the notion of *embedded multi-resolution mixtures* (EMM), which are families of embedded densities ranging over multiple image scales. EMMs are shown to generalize color histograms, complementing them with the ability to capture spatial image dependencies and allowing fine control over the invariance properties of the overall image representation. We present a cross-validation algorithm for finding the best multi-resolution decomposition, and determining the associated optimal subspace dimension, that is computationally efficient and has good retrieval performance.

Overall, the retrieval architecture composed by the Bayesian similarity criteria, a multi-resolution feature transformation, and the embedded mixture representation achieves a good compromise between retrieval accuracy, invariance, perceptual relevance of similarity judgments, and complexity. We illustrate these properties with an extensive experimental evaluation on three different databases that stress different aspects of the retrieval problem: the Brodatz texture database, the Columbia object database, and the Corel database of stock photography. In all cases, the new approach outperforms solutions representative of the state-of-the-art both in terms of objective (precision/recall) and subjective (perceptual) evaluation.

The paper is organized as follows. Section 2 establishes some notation to be used in the remaining sections. Section 3 introduces the decision-theoretic retrieval formulation, reviews some known results, and establishes the relationships between the Bayesian similarity criteria and various other similarity functions in common use in the literature. A theoretic characterization of the impact of the feature transformation and representation on the overall retrieval performance is then carried out on section 4. This characterization is also used to expose the major limitations of the color-based and texture-based retrieval strategies. Section 5 introduces embedded feature spaces and translates the theoretical results of Section 4 into a series of conditions for the design of the optimal feature transformation and representation. The embedded multi-resolution mixture representation is then introduced in section 6, which also provides an algorithm for optimal feature design. Finally, an experimental evaluation of the various aspects of decision-theoretic retrieval is presented in section 7.

2 Terms and notation

We start by introducing some notation. The basic element of image representation is an *image observation*. This can be a single pixel or a number d of them located in a pre-defined spatial neighborhood. We denote the space of observations by $\mathcal{Z} \subset \mathbb{R}^d$. The scalar d is always used to denote the dimension of the space $\mathcal{Z}$. Observations are mapped into feature vectors by a *linear* transformation

$$T : \mathcal{Z} \to \mathcal{X}. \tag{1}$$

We refer to $\mathcal{X}$ as the *feature space*, and $\mathbf{x} = T(\mathbf{z})$ a *feature vector*. *Features* are the elements of a feature vector. The matrix that defines the transformation T is denoted by $\mathbf{T}$.

A *feature representation* is a probabilistic model for how each of the image classes in the database populates the feature space $\mathcal{X}$. We introduce a *class indicator* variable $Y \in \{1, \ldots, M\}$ and denote the probability density function (pdf) of class i by $P_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = i)$. This also illustrates the following conventions for notation: random variables are represented in upper-case while values appear in lower-case, vectors are represented in boldface while scalars appear in normal type. Whenever the meaning is clear from context we replace the expression above by the simpler $P_{\mathbf{X}|Y}(\mathbf{x}|i)$. Finally, we frequently write $\mathbf{X} \in \mathcal{X}$ to indicate that the random variable $\mathbf{X}$ takes values in $\mathcal{X}$.

Throughout this work we assume that feature vectors are independent and identically distributed (iid), e.g.

$$P_{\mathbf{X}_1,\ldots,\mathbf{X}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{j=1}^{N} P_{\mathbf{X}}(\mathbf{x}_j).$$

One distribution that we will encounter frequently is the Gaussian, of mean μ and covariance $\mathbf{\Sigma}$,

$$P_{\mathbf{X}}(\mathbf{x}) = \mathcal{G}(\mathbf{x}, \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} e^{-\frac{1}{2}||\mathbf{x}-\mu||^2_{\mathbf{\Sigma}}} \tag{2}$$

where

$$||\mathbf{x} - \mu||_{\mathbf{\Sigma}} = \sqrt{(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)} \tag{3}$$

is the quadratic norm defined by $\mathbf{\Sigma}^{-1}$. The Euclidean norm is the particular case in which $\mathbf{\Sigma} = \mathbf{I}$. Another model that we will frequently refer to is the *histogram*. The histogram of a collection of feature vectors $\mathcal{S}$ is a vector $\mathbf{f} = \{f_1, \ldots, f_R\}$ associated with a partition of the feature space $\mathcal{X}$ into R regions $\{\mathcal{C}_1, \ldots, \mathcal{C}_R\}$, where f_r is the number of vectors in $\mathcal{S}$ landing on cell $\mathcal{C}_r$.

3 Decision-theoretic image similarity

In the CBIR context, image similarity can be formulated as a problem of statistical classification. Given the feature space $\mathcal{X}$, a retrieval system is simply a map

$$g : \mathcal{X} \to \{1, \ldots, M\}$$

from $\mathcal{X}$ to the index set of the M classes in the database. In this sense, it is natural to adopt a decision-theoretic formulation of the retrieval problem, where the goal is to design systems that have *minimum probability of retrieval error*, i.e. that are wrong as rarely as possible.

Definition 1 *A minimum probability of error (MPE) retrieval system is the mapping*

$$g : \mathcal{X} \to \{1, \ldots, M\}$$

that minimizes

$$P_{\mathbf{X},Y}(g(\mathbf{X}) \neq Y).$$

Under this definition, the optimal similarity function is well known [15].

Theorem 1 *Given a feature space $\mathcal{X}$ and a query $\mathbf{x}$, the similarity function that minimizes the probability of retrieval error is the Bayes or maximum a posteriori (MAP) classifier*

$$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}). \tag{4}$$

Furthermore, the probability of error is lower bounded by the Bayes error

$$L^*_{\mathcal{X}} = 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \tag{5}$$

where $E_{\mathbf{x}}$ means expectation with respect to $P_{\mathbf{X}}(\mathbf{x})$.

Proof: See appendix A.1.

One way to implement the MAP classifier is with recourse to Bayes rule

$$\begin{aligned} g^*(\mathbf{x}) &= \arg\max_i \prod_{j=1}^{N} P_{\mathbf{X}|Y}(\mathbf{x}_j|i) P_Y(y=i) \\ &= \arg\max_i \sum_{j=1}^{N} \log P_{\mathbf{X}|Y}(\mathbf{x}_j|i) + \log P_Y(i), \end{aligned} \tag{6}$$

where we have used the iid assumption for $\mathbf{X}$. Equation (6) is denoted by *Bayesian retrieval criteria* and image retrieval based on it as *decision-theoretic retrieval* (DTR).

3.1 A unified view of image similarity

In this section, we analyze the relationships between the Bayesian retrieval criteria and a significant number of previously proposed similarity functions. The goal is to show that many of the latter can be derived from the decision-theoretic principles at the core of the former, by making various assumptions or approximations. This not only demonstrates that, in general, these alternatives cannot lead to superior performance but also enables a principled understanding of their limitations and applicability to different retrieval contexts.

The assumptions/approximations required to derive several popular similarity functions from the Bayesian criteria are depicted in Figure 1. If an upper bound on the Bayes error of a collection of two-way classification problems is minimized instead of the probability of error of the original problem, the Bayesian criteria reduces to the *Bhattacharyya distance* (BD). On the other hand, if the original criteria is minimized, but the different image classes are assumed to be equally likely a priori, we have the *maximum likelihood* (ML) retrieval criteria. As the number of query vectors grows to infinity the ML criteria tends to the *minimum discrimination information* (MDI), which in turn can be approximated by the χ^2 test by performing a simple first order Taylor series expansion. Alternatively, MDI can be simplified by assuming that the underlying probability densities belong to a pre-defined family. For *auto-regressive sources* it reduces to the *Itakura-Saito* distance that has received significant attention in the speech literature. In the Gaussian case, further assumption of orthonormal covariance matrices leads to the *quadratic distance* (QD) frequently found in the compression literature. The next possible simplification is to assume that all classes share the same covariance matrix, leading to the MD. Finally, assuming identity covariances results in the square of the *Euclidean distance* (ED). We next derive in more detail all these relationships.

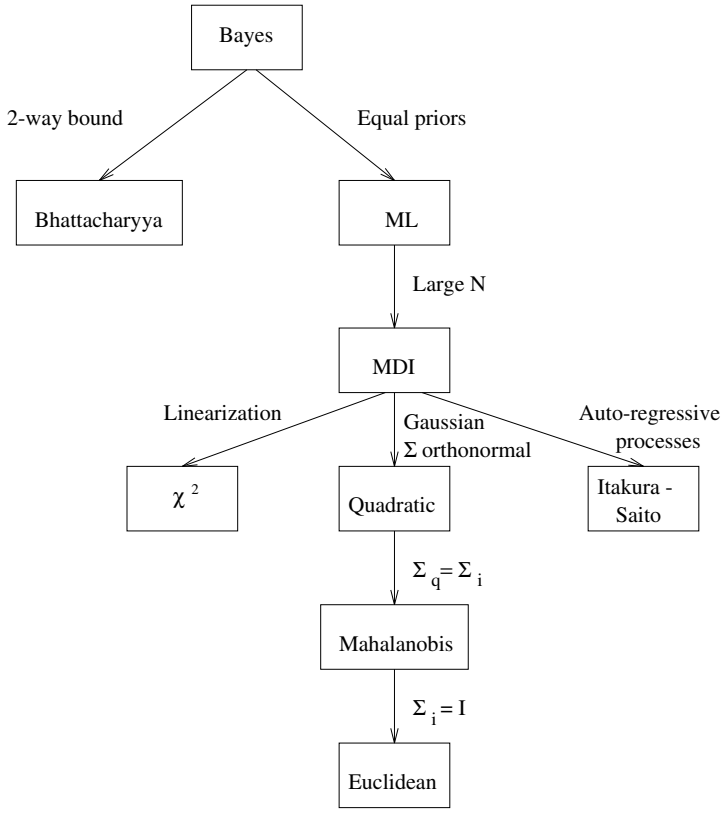

Figure 1: Relations between different image similarity functions.

3.1.1 Bhattacharyya distance

If there are only two classes in the classification problem, (5) can be written as [22]

$$\begin{aligned} L^*_{\mathcal{X}} &= E_{\mathbf{x}}[\min(P_{Y|\mathbf{X}}(0|\mathbf{x}), P_{Y|\mathbf{X}}(1|\mathbf{x}))] \\ &= \int P_{\mathbf{X}}(\mathbf{x}) \min[P_{Y|\mathbf{X}}(0|\mathbf{x}), P_{Y|\mathbf{X}}(1|\mathbf{x})] d\mathbf{x} \\ &= \int \min[P_{\mathbf{X}|Y}(\mathbf{x}|0)P_Y(0), P_{\mathbf{X}|Y}(\mathbf{x}|1)P_Y(1)] d\mathbf{x} \\ &\leq \sqrt{P_Y(0)P_Y(1)} \int \sqrt{P_{\mathbf{X}|Y}(\mathbf{x}|0)P_{\mathbf{X}|Y}(\mathbf{x}|1)} d\mathbf{x}, \end{aligned}$$

where we have used the bound $\min[a, b] \leq \sqrt{ab}$. The last integral is usually known as the Bhattacharyya distance between $P_{\mathbf{X}|Y}(\mathbf{x}|0)$ and $P_{\mathbf{X}|Y}(\mathbf{x}|1)$ and has been proposed (e.g. [47, 11]) for image retrieval where, for a query density $P_{\mathbf{X}}(\mathbf{x})$, it takes the form

$$g(\mathbf{x}) = \arg\min_i \int \sqrt{P_{\mathbf{X}}(\mathbf{x}) P_{\mathbf{X}|Y}(\mathbf{x}|i)} d\mathbf{x}. \tag{7}$$

The resulting classifier can thus be seen as the one which finds the lowest upper-bound on the Bayes error for the collection of two-class problems involving the query and each of the database classes.

Whenever it is possible to solve the minimization of the error probability on the multi-class retrieval problem it makes small sense to replace it by the search for the two class problem with the smallest error bound. Consequently, the above interpretation of the BD makes it clear that, in general, there is small justification to prefer it to DTR.

3.1.2 Maximum likelihood

It is a straightforward consequence of (6) that, when all image classes are a priori equally likely, $P_Y(i) = 1/M$,

$$g(\mathbf{x}) = \arg\max_i \frac{1}{N} \sum_{j=1}^{N} \log P_{\mathbf{X}|Y}(\mathbf{x}_j|i). \tag{8}$$

This decision rule is known as the maximum likelihood classifier. While class priors $P_Y(i)$ can provide a useful mechanism to 1) account for the context in which the retrieval operation takes place, 2) integrate information from multiple content modalities that may be available in the database, and 3) design learning algorithms [81, 79], in this work we assume that there is no a priori reason to prefer any given image over the rest. In this case, Bayesian and maximum likelihood retrieval are equivalent.

3.1.3 Minimum discrimination information

If $H_i, i = 1, 2$, are the hypotheses that $\mathbf{x}$ is drawn from the statistical population with density $P_i(\mathbf{x})$, the *Kullback-Leibler divergence* (KLD) or *relative entropy* [33, 13]

$$KL[P_2(\mathbf{x})||P_1(\mathbf{x})] = \int P_2(\mathbf{x}) \log \frac{P_2(\mathbf{x})}{P_1(\mathbf{x})} d\mathbf{x} \tag{9}$$

measures the mean information per observation from $P_2(\mathbf{x})$ for discrimination in favor of H_2 against H_1. Because it measures the difficulty of discriminating between the two populations, and is always non-negative and equal to zero only when $P_1(\mathbf{x}) = P_2(\mathbf{x})$ [33], the KLD has been proposed as a measure of similarity for various compression and signal processing problems [27, 36, 18, 10].

Given a density $P_1(\mathbf{x})$ and a family of densities $\mathcal{M}$ the MDI criteria [33] seeks the density in $\mathcal{M}$ that is the "nearest neighbor" of $P_1(\mathbf{x})$ in the KLD sense

$$P_2^*(\mathbf{x}) = \arg \min_{P_2(\mathbf{x}) \in \mathcal{M}} KL[P_2(\mathbf{x})||P_1(\mathbf{x})].$$

If $\mathcal{M}$ is a large family, containing $P_1(\mathbf{x})$, this problem has the trivial solution $P_2(\mathbf{x}) = P_1(\mathbf{x})$, which is not always the most interesting. In other cases, a sample from $P_2(\mathbf{x})$ is available but the explicit form of the distribution is not known. In these situations it may be more useful to seek for the distribution that minimizes the KLD subject to a stricter set of constraints. Kullback suggested the problem

$$P_2^*(\mathbf{x}) = \arg \min_{P_2(\mathbf{x}) \in \mathcal{M}} KL[P_2(\mathbf{x})||P_1(\mathbf{x})]$$

subject to

$$\int T(\mathbf{x}) P_2(\mathbf{x}) = \theta$$

where $T(\mathbf{x})$ is a measurable statistic (e.g. the mean when $T(\mathbf{x}) = \mathbf{x}$) and θ can be computed from a sample (e.g. the sample mean). He showed that the minimum is 1) achieved by

$$P_2^*(\mathbf{x}) = \frac{1}{Z} e^{-\lambda T(\mathbf{x})} P_1(\mathbf{x})$$

where Z is a normalizing constant, $Z = \int e^{-\lambda T(\mathbf{x})} P_1(\mathbf{x}) d\mathbf{x}$, and λ a Lagrange multiplier that weighs the importance of the constraint; and 2) equal to

$$KL[P_2^*(\mathbf{x})||P_1(\mathbf{x})] = -\lambda\theta - \log Z.$$

Gray and his colleagues have studied extensively the case in which $P_1(\mathbf{x})$ belongs to the family of *auto-regressive moving average* (ARMA) processes [27, 19] and showed, among other things, that in this case the optimal solution is a variation of the Itakura-Saito distance commonly used in speech analysis and compression. Kupperman [34] has shown that when all densities are members of the exponential family, the constrained version of MDI is equivalent to maximum likelihood.

The KLD has only been recently considered in the retrieval literature [78, 31, 56, 7, 16], where attention has focused on the unconstrained MDI problem

$$g(\mathbf{x}) = \arg\min_i KL[P_{\mathbf{X}}(\mathbf{x})||P_{\mathbf{X}|Y}(\mathbf{x}|i)], \tag{10}$$

where $P_{\mathbf{X}}(\mathbf{x})$ is the density of the query and $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ that of the i^{th} image class. Similarly to the constrained case, it is possible to derive a connection between unconstrained MDI and maximum likelihood. However, the connection is much stronger in the unconstrained case since there is no need to make any assumptions regarding the type of densities involved. In particular, by simple application of the law of large numbers to (8),

$$\begin{aligned}
g(\mathbf{x}) &= \arg\max_i E_{\mathbf{x}}[\log P_{\mathbf{X}|Y}(\mathbf{x}|i)] \text{ as } N \to \infty \\
&= \arg\max_i \int P_{\mathbf{X}}(\mathbf{x}) \log P_{\mathbf{X}|Y}(\mathbf{x}|i) d\mathbf{x} \\
&= \arg\min_i \int P_{\mathbf{X}}(\mathbf{x}) \log P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} - \int P_{\mathbf{X}}(\mathbf{x}) \log P_{\mathbf{X}|Y}(\mathbf{x}|i) d\mathbf{x} \\
&= \arg\min_i \int P_{\mathbf{X}}(\mathbf{x}) \log \frac{P_{\mathbf{X}}(\mathbf{x})}{P_{\mathbf{X}|Y}(\mathbf{x}|i)} d\mathbf{x} \\
&= \arg\min_i KL[P_{\mathbf{X}}(\mathbf{x})||P_{\mathbf{X}|Y}(\mathbf{x}|i)],
\end{aligned}$$

where $E_{\mathbf{x}}$ is the expectation with respect to the query density $P_{\mathbf{X}}(\mathbf{x})$. This means that, independently of the type of densities, MDI is simply the asymptotic limit of the ML criteria as the cardinality of the query grows to infinity. This relationship is important for various reasons. First, it confirms that the Bayesian criteria converges to a meaningful similarity function between image densities as the cardinality of the query grows. Second, it makes it clear that while ML and MDI perform equally well for image-based queries, the Bayesian criteria has the added advantage of also enabling queries based on image regions. Finally, it establishes a connection between the Bayesian criteria and several similarity functions that can be derived from MDI.

3.1.4 χ^2 test

The first of such similarity functions is the χ^2 statistic. Using a first order Taylor series approximation for the logarithmic function about $x = 1$, $\log(x) \approx x - 1$, we obtain

$$\begin{aligned}
KL[P_1(\mathbf{x})||P_2(\mathbf{x})] &= \int P_1(\mathbf{x}) \log \frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} d\mathbf{x} \\
&\approx \int \frac{P_1(\mathbf{x})^2 - P_1(\mathbf{x}) P_2(\mathbf{x})}{P_2(\mathbf{x})} d\mathbf{x}
\end{aligned}$$

$$= \int \left(\frac{P_1(\mathbf{x})^2 - P_1(\mathbf{x})P_2(\mathbf{x})}{P_2(\mathbf{x})} - P_1(\mathbf{x}) + P_2(\mathbf{x}) \right) d\mathbf{x}$$

$$= \int \frac{(P_1(\mathbf{x}) - P_2(\mathbf{x}))^2}{P_2(\mathbf{x})} d\mathbf{x},$$

where we have used the fact that $\int P_i(\mathbf{x})d\mathbf{x} = 1, i = 1, 2$. In the retrieval context, this means that MDI can be approximated by

$$g(\mathbf{x}) \approx \arg\min_i \int \frac{(P_{\mathbf{X}}(\mathbf{x}) - P_{\mathbf{X}|Y}(\mathbf{x}|i))^2}{P_{\mathbf{X}|Y}(\mathbf{x}|i)} d\mathbf{x}. \tag{11}$$

The integral on the right is known as the χ^2 statistic and the resulting criteria a χ^2 test [51]. It has been proposed as a metric for image similarity in [61, 7, 56, 35], among others. Since it results from the linearization of the KLD, it can be seen as an approximation to the asymptotic limit of the ML criteria. Obviously, this linearization can discard a significant amount of information and there is, in general, no reason to believe that it should perform better than DTR.

3.1.5 The Gaussian case

Several similarity functions of practical interest can be derived from the Bayesian criteria when the class likelihood functions are Gaussian. In this case, (8) becomes

$$g(\mathbf{x}) = \arg\min_i \log|\mathbf{\Sigma}_i| + \frac{1}{N}\sum_n (\mathbf{x}_n - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x}_n - \mu_i)$$

$$= \arg\min_i \log|\mathbf{\Sigma}_i| + \hat{\mathcal{L}}_i, \tag{12}$$

where

$$\hat{\mathcal{L}}_i = \frac{1}{N}\sum_n (\mathbf{x}_n - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x}_n - \mu_i)$$

is the *quadratic distance* (QD) commonly found in the perceptually weighted compression literature [24, 38]. As a retrieval metric, the QD can thus be seen as the result of imposing two stringent restrictions on the generic ML criteria. First, that all image sources are Gaussian and, second, that their covariance matrices are orthonormal ($|\mathbf{\Sigma}_i| = 1, \forall i$). Furthermore, because

$$\hat{\mathcal{L}}_i = \frac{1}{N}\sum_n (\mathbf{x}_n - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x}_n - \mu_i)$$

$$= \frac{1}{N}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}} + \hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x}_n - \hat{\mathbf{x}} + \hat{\mathbf{x}} - \mu_i)$$

$$= \frac{1}{N} trace[\mathbf{\Sigma}_i^{-1} \sum_n (\mathbf{x}_n - \hat{\mathbf{x}})(\mathbf{x}_n - \hat{\mathbf{x}})^T] + (\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\hat{\mathbf{x}} - \mu_i)^T$$

$$= \quad trace[\boldsymbol{\Sigma}_i^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}] + \mathcal{M}_i, \tag{13}$$

where $\hat{\mathbf{x}} = 1/N \sum_n \mathbf{x}_n$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}} = 1/N \sum_n (\mathbf{x}_n - \hat{\mathbf{x}})(\mathbf{x}_n - \hat{\mathbf{x}})^T$ are, respectively, the sample mean and covariance of $\mathbf{x}_n$ and

$$\mathcal{M}_i = (\hat{\mathbf{x}} - \mu_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\mathbf{x}} - \mu_i)^T$$

the Mahalanobis distance, we see that the MD results from complementing Gaussianity with the assumption that all classes have the same covariance ($\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}, \forall i$).

Finally, if this covariance is the identity ($\boldsymbol{\Sigma} = \mathbf{I}$), we obtain the square of the Euclidean distance (ED) or *mean squared error*

$$\mathcal{E}_i = (\hat{\mathbf{x}} - \mu_i)^T (\hat{\mathbf{x}} - \mu_i). \tag{14}$$

The MD, the ED, and variations on both, have been widely used in the retrieval literature [64, 41, 1, 65, 49, 62, 54, 44, 53, 6, 56, 70, 59, 3].

3.1.6 Some intuition for the advantages of DTR

The analysis of the Gaussian case emphasizes why there is little justification to prefer any of the above three similarity metrics to the Bayesian criteria. Recall that while for the latter the similarity function is

$$g(\mathbf{x}) = \arg\min_i \log|\boldsymbol{\Sigma}_i| + \overbrace{trace[\boldsymbol{\Sigma}_i^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}] + \underbrace{(\hat{\mathbf{x}} - \mu_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\mathbf{x}} - \mu_i)^T}_{\text{MD}}}^{\text{QD}}, \tag{15}$$

all other three are approximations that arbitrarily discard covariance information.

As shown in Figure 2, this information is important for the detection of subtle variations such as rotation and scaling in feature space. In a) and b), we show the distance, under both QD and MD between a Gaussian and a replica rotated by $\theta \in [0, \pi]$. Plot b) clearly illustrates that while the MD has no ability to distinguish between the rotated Gaussians, the inclusion of the $trace[\boldsymbol{\Sigma}_i^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}]$ term leads to a much more intuitive measure of similarity: minimum when both Gaussians are aligned and maximum when they are rotated by $\pi/2$.

As illustrated by c) and d), further inclusion of the term $\log|\boldsymbol{\Sigma}_i|$ (full ML retrieval) penalizes mismatches in scaling. In plot c), we show two Gaussians, with covariances $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{I}$ and $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}$, centered on zero. In this example, MD is always zero, while $trace[\boldsymbol{\Sigma}_i^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}] \propto 1/\sigma^2$ penalizes small σ and $\log|\boldsymbol{\Sigma}_i| \propto \log\sigma^2$ penalizes large σ. The total distance is shown as a function of $\log\sigma^2$ in plot d) where, once again, we observe an intuitive behavior: the penalty is minimal when both Gaussians have the same scale ($\log\sigma^2 = 0$), increasing monotonically with the amount of scale mismatch. Notice that if the $\log|\boldsymbol{\Sigma}_i|$ term is not included, large changes in scale may not be penalized at all.

3.1.7 L^p norms

Despite all its good properties, the Bayesian retrieval criteria has received small attention in the context of CBIR. An overwhelmingly more popular similarity function is

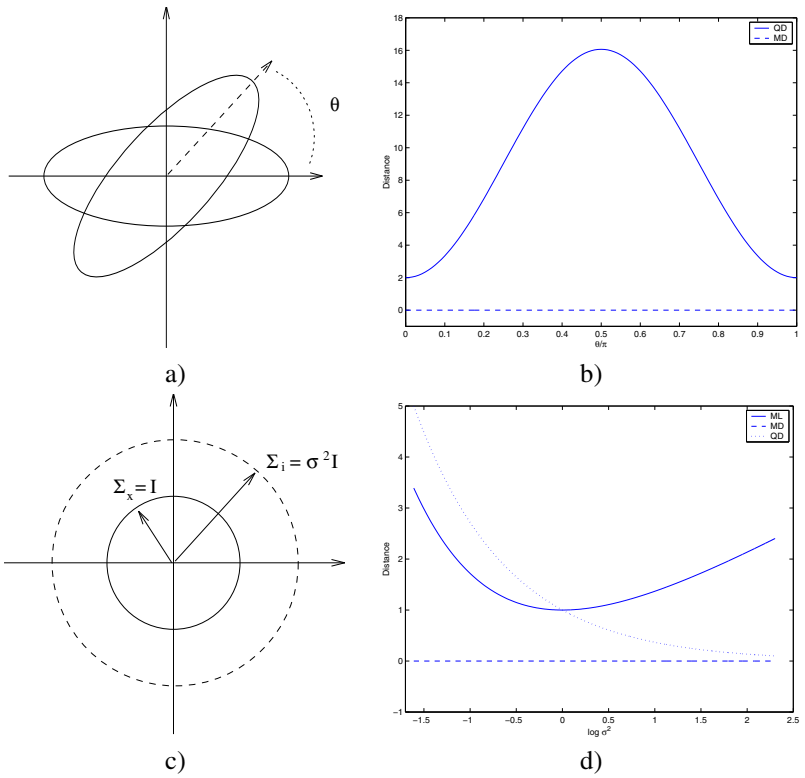

Figure 2: a) A Gaussian with mean $(0,0)^T$ and covariance $diag(4, 0.25)$ and its replica rotated by θ. b) Distance between the Gaussian and its rotated replica as a function of θ/π under both the QD and the MD. c) Two Gaussians with different scales. d) Distance between them as a function of $\log \sigma^2$ under ML, QD, and MD.

the L^p norm of the difference between densities

$$g(\mathbf{X}) = \arg\min_i \left(\int_{\mathcal{X}} |P_{\mathbf{X}}(\mathbf{x}) - P_{\mathbf{X}|Y}(\mathbf{x}|i)|^p d\mathbf{x} \right)^{\frac{1}{p}}. \quad (16)$$

These norms are particularly common in the color-based retrieval literature as similarity metrics for color histograms. Defining $\mathbf{q}$ to be the histogram of Q query vectors, and $\mathbf{p}^i$ the histogram of P^i vectors from the i^{th} image class, (16) reduces to

$$g(\mathbf{X}) = \arg\min_i \left(\sum_r \left| \frac{q_r}{Q} - \frac{p_r^i}{P^i} \right|^p \right)^{\frac{1}{p}},$$

As shown in [68], when the histograms are normalized ($\sum_r q_r/Q = \sum_r p_r^i/P^i = 1, \forall i$), the minimization of the L^1 distance is equivalent to the maximization of the HI

$$g(\mathbf{X}) = \arg\max_i \frac{\sum_r \min(q_r, p_r^i)}{Q}. \quad (17)$$

While (8) minimizes the classification error, (16) implies that minimizing pointwise similarity between density estimates should be the ultimate retrieval criteria. Clearly, for any of the two criteria to work, it is necessary that the estimates be close to the true densities. However, it is known (e.g. see Theorem 6.5 of [15]) that the probability of error of rules of the type of (8) tends to the Bayes error orders of magnitude faster than the associated density estimates tend to the right distributions. This implies that accurate density estimates are not required everywhere for the classification criteria to work.

In fact, accuracy is required only in the regions near the boundaries between the different classes, because these are the only regions that matter for the classification decisions. On the other hand, the criteria of (16) is clearly dependent on the quality of the density estimates all over $\mathcal{X}$. It, therefore, places a much more stringent requirement on the quality of these estimates and, since density estimation is know to be a difficult problem [76, 63], it is unlikely that it will perform better than (8). This is indeed confirmed by the experimental results presented in section 7.

4 Decision-theoretic guidelines for image representation

One of the interesting properties of the DTR formulation is that it enables the design of systems where all components are optimized with respect to a common criteria (probability of retrieval error). We next analyze how the feature transformation and representation impact the overall system optimality.

4.1 Feature transformation

We start by analyzing the role of the feature transformation.

Theorem 2 *Given a retrieval system with observation space $\mathcal{Z}$ and a feature transformation*

$$T : \mathcal{Z} \to \mathcal{X},$$

then

$$L^*_{\mathcal{X}} \geq L^*_{\mathcal{Z}} \tag{18}$$

where $L^*_{\mathcal{Z}}$ and $L^*_{\mathcal{X}}$ are, respectively, the Bayes errors on $\mathcal{Z}$ and $\mathcal{X}$. Furthermore, equality is achieved if and only if T is an invertible transformation.

Proof: see appendix A.2.

The last statement of the theorem is a worst-case result. In fact, for a specific retrieval problem, it may be possible to find non-invertible feature transformations that do not increase Bayes error. What is not possible is to find 1) a feature transformation that will reduce the Bayes error, or 2) a universal non-invertible feature transformation guaranteed not to increase the Bayes error on all retrieval problems.

4.2 Feature representation

While a necessary condition, low Bayes error is not sufficient for accurate retrieval since the actual error may be much larger than the lower bound.

Theorem 3 *Given a retrieval system with a feature space $\mathcal{X}$, unknown class probabilities $P_Y(i)$, class densities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, and a decision function*

$$g(\mathbf{x}) = \arg\max_i \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i), \tag{19}$$

the actual probability of error is upper bounded by

$$P(g(\mathbf{X}) \neq Y) \leq L^*_{\mathcal{X}} + \sum_i \int |P_{\mathbf{X}|Y}(\mathbf{x}|i)P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i)|d\mathbf{x}. \tag{20}$$

Proof: see appendix A.3.

In the remainder of this work we assume that the classes are a-priori equiprobable, i.e. $P_Y(i) = 1/M, \forall i$. This leads to the following corollary.

Corollary 1 *Given a retrieval problem with equiprobable classes, a feature space $\mathcal{X}$, unknown class conditional likelihood functions $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, and a decision function*

$$g(\mathbf{x}) = \arg\max_i \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i), \tag{21}$$

the difference between the actual and Bayes error is upper bounded by

$$P(g(\mathbf{X}) \neq Y) - L^*_{\mathcal{X}} \leq \Delta_{g,\mathcal{X}} \tag{22}$$

where

$$\Delta_{g,\mathcal{X}} = \sum_i KL[P_{\mathbf{X}|Y}(\mathbf{x}|i)||\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)], \tag{23}$$

is the estimation error.

Proof: see Appendix A.4.

In summary, the difference between the actual probability of retrieval error and the Bayes error is upper bounded by the error in the density estimates. This implies that, if the Bayes error is small, accurate density estimation is a sufficient condition for high retrieval accuracy. In particular, good density estimation will suffice to guarantee optimal performance when the feature transformation is invertible.

4.3 Strategies for image representation

The two theorems are convenient tools for analyzing the balance between feature transformation and representation achieved by any retrieval strategy. We now proceed to do so for two strategies in widespread use in the literature.

4.3.1 The color strategy

The theorems suggest that all that really matters for accurate retrieval is good density estimation. Since no feature transformation can reduce the Bayes error, there seems to be no advantage in using one. This is the rationale behind Strategy 1 (S1): *avoid feature transformations altogether and do all the estimation directly in* $\mathcal{Z}$. As Figure 3 illustrates, the main problem with this strategy is that density estimation can be difficult in $\mathcal{Z}$. Significant emphasis must therefore be given to the feature representation which is required to rely on a sophisticated density model. One possible solution, that has become a de-facto standard for color-based retrieval [68, 56, 58, 41, 59, 64, 66, 1], is the histogram. This solution is illustrated in Figure 3 b).

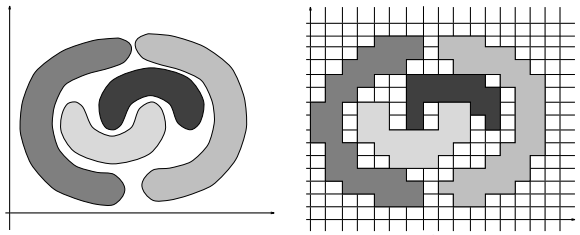

Figure 3: Example of a retrieval problem with four image classes. a) In the space of image observations, the class densities can have complicated shapes. b) Strategy 1 is to simply model the class densities as accurately as possible.

While they work reasonably well when $\mathcal{Z}$ is a low-dimensional space, e.g. the 3-D space of pixel colors, histograms are of very limited use in high dimensions. This is a consequence of the exponential growth of the number of histogram cells with the dimension of the space. Since this dimension is proportional to the size of the region of support of the observations, accurate histogram-based density estimates can only be obtained for very small spatial neighborhoods. Consequently, *the representation*

cannot capture the spatial dependencies that are crucial for fine image discrimination. This is illustrated by Figure 4.

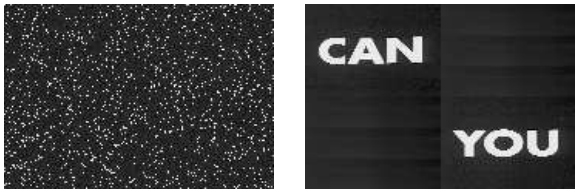

Figure 4: Two images that, although visually very dissimilar, have the same color histogram.

4.3.2 The texture strategy

Because accurate density estimation is usually a difficult problem, a feature transformation can be helpful if it makes estimation significantly easier in $\mathcal{X}$ than what it is in $\mathcal{Z}$. The rationale behind Strategy 2 (S2) is to exploit this as much as possible: *find a feature transformation that clearly separates the image classes in $\mathcal{X}$, rendering estimation trivial*. Ideally, in $\mathcal{X}$, each class should be characterized by a simple parametric density, such as the Gaussians in Figure 5, and a simple classifier should be able to guarantee performance close to the Bayes error.

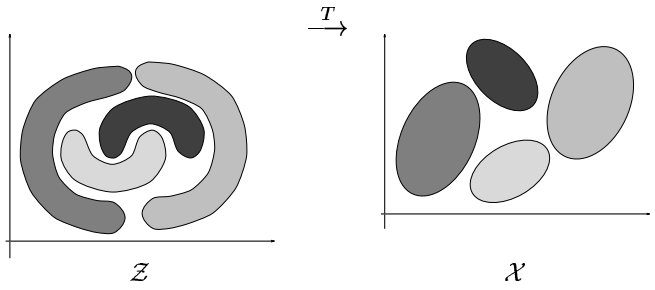

Figure 5: Example retrieval problem with four image classes. Strategy 2 is to find a feature transformation such that density estimation is much easier in $\mathcal{X}$ than in $\mathcal{Z}$.

Strategy S2 has become prevalent in the texture literature, where numerous feature transformations have been proposed to achieve *good discrimination* between different texture classes [64, 41, 54, 44, 55, 17, 45, 73, 57, 69, 9, 71]. These transformations are then combined with simple similarity functions, like the Mahalanobis and Euclidean distances or variations of these, that assume Gaussianity in $\mathcal{X}$. More recently it has also been embraced by many retrieval systems [6, 49, 70, 59, 56, 64, 41, 53, 3].

The main problem of strategy S2 is the assumption that it is always possible to find a transformation that maps a collection of complicated densities in $\mathcal{Z}$ into a collection

of Gaussians in $\mathcal{X}$, *without compromising Bayes error*. This is usually not possible and, for acceptable levels of Bayes error, feature densities tend to have non-trivial shapes, e.g. are multi-modal. Assuming a Gaussian model on $\mathcal{X}$ can therefore lead to poor density estimates and significant penalties in retrieval precision.

5 Optimal retrieval systems

The two standard strategies can be seen as two ends of a continuum: while S1 is intransigent with respect to any loss in Bayes error and therefore asks too much of the feature representation; S2 constrains the representation to trivial models, expecting the feature transformation to do the impossible. It seems that a wiser position would be to stand somewhere in between. Since the overall probability of error is upper bounded by the sum of the Bayes and estimation errors, we need to consider the two *simultaneously*.

5.1 Optimal feature representation

With respect to the feature representation, Theorem 3 shows that (whatever the feature transformation may be) there are no guarantees of small probability of error if the density estimates are inaccurate. The quality of these estimates is determined by two factors: the choice of probabilistic model, or feature representation, and the estimation of the parameters of this model.

5.1.1 Parameter estimation

The following lemma shows that, for a given parametric density family, the optimal parameters are obtained by standard maximum likelihood estimation.

Lemma 1 *Consider a retrieval problem with equiprobable classes, a feature space $\mathcal{X}$, a decision function*

$$g(\mathbf{x};\mathbf{p}) = \arg\max_i P_{\mathbf{X}|Y}(\mathbf{x}|i;\mathbf{p}_i), \tag{24}$$

where $P_{\mathbf{X}|Y}(\mathbf{x}|i;\mathbf{p}_i)$ is a pdf parameterized by $\mathbf{p}_i$, and a sequence of samples $\{\mathbf{x}_i, i \in 1,\ldots,M\}$, where $\mathbf{x}_i$ contains N iid feature vectors from the i^{th} image class in the database. Then the upper bound on the density estimation error

$$\Delta_{g,\mathcal{X}} = \sum_i KL[P_{\mathbf{X}|Y}(\mathbf{x}|i)||P_{\mathbf{X}|Y}(\mathbf{x}|i;\mathbf{p}_i)], \tag{25}$$

is minimized if

$$\mathbf{p}_i = \arg\max_{\mathbf{p}} \frac{1}{N}\sum_{k=1}^{N} \log P_{\mathbf{X}|Y}(\mathbf{x}_{i,k}|i;\mathbf{p}). \tag{26}$$

Proof: see Appendix A.5

The main difficulty posed by feature representation is, therefore, to determine what is the best parametric family for density estimation.

5.1.2 Parametric density families

We have seen in the previous section that the Gaussian and histogram models can impose strong limitations on retrieval accuracy. There are, however, several more sophisticated density models including vector quantizers [25], decision-trees [8], mixtures [72], and kernel-based representations [63]. While all of the latter overcome the main limitations of the former, they introduce some problems of their own.

For example, kernel-based density estimates do not provide a compact description of the underlying density (their complexity is proportional to the number of feature vectors in the training set) and lead to a similarity function (8) that is too complex for most retrieval applications. On the other hand, vector quantizers and decision-trees assume a partition of the feature space into mutually exclusive cells that can originate significant fluctuations of the density estimates in the presence of small variations of the true density [80]. In fact, these representations can be seen as generalizations of the histogram that, while overcoming the problem of exponential complexity in the dimension of the space, still exhibit all the limitations associated with a partition of the feature space into non-overlapping cells. Such limitations are avoided by mixture models.

Definition 2 *A mixture density is a density of the form*

$$P_{\mathbf{X}}(\mathbf{x}) = \sum_{w=1}^{C} P_{\mathbf{X}|W}(\mathbf{x}|w) P_W(w), \tag{27}$$

where $\{P_{\mathbf{X}|W}(\mathbf{x}|w)\}_{w=1}^{C}$ *is a sequence of mixture components.*

Mixture models are particularly well suited for the retrieval problem due to four main properties. First, because the mixture inherits the complexity of its components, it is tractable in high dimensions whenever the components are. In the Gaussian case, complexity is only quadratic in the dimension of the space (linear for Gaussians of diagonal covariance). Mixtures are therefore significantly more tractable than histograms. Second, like histograms, mixtures can approximate arbitrary densities. In fact, because they rely on smoother kernels, approximations based on mixtures can be significantly better than those possible with histograms, vector quantizers, or decision trees [63, 39, 37]. Third, as is clear from (27), the complexity of a mixture is linear in the number of components C, which is usually small. Hence, unlike kernel-based methods, mixtures provide a compact representation of the underlying density.

In this sense, mixtures combine the good properties of the Gaussian, histogram, and kernel-based models: computational tractability, smoothness, and expressiveness. A fourth property, which is particularly relevant in the context of this work, is that once a set of parameter estimates is available for a density defined on $\mathcal{X}$, the corresponding parameters on a sequence of important subspaces are automatically determined. We will return to this issue in section 6.

5.2 Optimal feature transformations

Unlike the feature representation, which affects only the estimation error, the choice of feature transformation has impact in both the Bayes and estimation errors. While the

impact on the Bayes error is direct (the Bayes error depends uniquely on the feature transformation), the impact on the estimation error is more subtle. It derives from the phenomena known as the curse of dimensionality: for a given amount of training data, the quality of density estimates degrades as the dimension of the feature space increases. The design of an optimal feature transformation must, therefore, account for both the Bayes and estimation errors. To understand the associated trade-offs we introduce the notion of embedded feature spaces.

5.2.1 Embedded feature spaces

Definition 3 *Given two vector spaces $\mathcal{X}_m$ and $\mathcal{X}_n$, $m < n$, such that $dim(\mathcal{X}_m) = m$ and $dim(\mathcal{X}_n) = n$ an embedding is a mapping*

$$\epsilon : \mathcal{X}_m \to \mathcal{X}_n \tag{28}$$

which is one-to-one.

A canonical example of embedding is the zero padding operator for Euclidean spaces

$$\iota_m^n : \mathbb{R}^m \to \mathbb{R}^n \tag{29}$$

where $\iota_m^n(\mathbf{x}) = (\mathbf{x}, \mathbf{0})$, $\mathbf{x} \in \mathbb{R}^m$, and $\mathbf{0} \in \mathbb{R}^{n-m}$.

Definition 4 *A sequence of vector spaces $\{\mathcal{X}_1, \ldots, \mathcal{X}_d\}$, such that $dim(\mathcal{X}_i) < dim(\mathcal{X}_{i+1})$, is called embedded if there exists a sequence of embeddings*

$$\epsilon_i : \mathcal{X}_i \to \mathcal{X}'_{i+1},\ i = 1, \ldots, d-1, \tag{30}$$

such that $\mathcal{X}'_{i+1} \subset \mathcal{X}_{i+1}$.

The inverse operation of an embedding is a submersion.

Definition 5 *Given two vector spaces $\mathcal{X}_m$ and $\mathcal{X}_n$, $m < n$, such that $dim(\mathcal{X}_m) = m$ and $dim(\mathcal{X}_n) = n$ a submersion is a mapping*

$$\gamma : \mathcal{X}_n \to \mathcal{X}_m \tag{31}$$

which is surjective.

A canonical example of submersion is the projection of Euclidean spaces along the coordinate axes

$$\pi_m^n : \mathbb{R}^n \to \mathbb{R}^m \tag{32}$$

where $\pi_m^n(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n) = (x_1, \ldots, x_m)$.

The following theorem shows that any linear feature transformation originates a sequence of embedded vector spaces with monotonically decreasing Bayes error, and monotonically increasing estimation error.

Theorem 4 *Let*

$$T : \mathbb{R}^d \to \mathcal{X} \subset \mathbb{R}^d,$$

be a linear feature transformation. Then,

$$\mathcal{X}_i = \pi_i^d(\mathcal{X}), i = 1, \ldots, d-1 \tag{33}$$

is a sequence of embedded feature spaces such that

$$L^*_{\mathcal{X}_{i+1}} \leq L^*_{\mathcal{X}_i}. \tag{34}$$

Furthermore, if $\mathbf{X}_1^d = \{\mathbf{X}_1, \ldots, \mathbf{X}_d\}$ *is a sequence or random variables such that* $\mathbf{X}_i \in \mathcal{X}_i$*,*

$$\mathbf{X}_i = \pi_i^d(\mathbf{X}), i = 1, \ldots, d \tag{35}$$

and $\{g(\mathbf{x})\}_1^d$ *a sequence of decision functions*

$$g_i(\mathbf{x}) = \arg\max_k \hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}|k) \tag{36}$$

then

$$\Delta_{g_{i+1},\mathcal{X}_{i+1}} \geq \Delta_{g_i,\mathcal{X}_i}. \tag{37}$$

Proof: see Appendix A.6.

It follows that, in general, it is impossible to minimize the Bayes and estimation errors simultaneously. On one hand, given a feature space $\mathcal{X}_i$ it is usually possible to find a subspace where density estimates are more accurate. On the other, the projection onto this subspace will increase the Bayes error. The practical result is that there is always a need to reach a compromise between the two sources of error. This is illustrated by Figure 6 which shows the typical evolution of the upper and lower bounds on the probability of error as one considers successively higher-dimensional subspaces of a feature space $\mathcal{X}$.

Since accurate density estimates can usually be obtained in low-dimensional spaces, the two bounds tend to be close when the subspace dimension is small. In this case, the probability of error is dominated by the Bayes error. For higher-dimensional subspaces the decrease in Bayes error is canceled by an increase in estimation error and the actual probability of error increases. Overall, the curve of the probability of error exhibits the convex shape depicted in the figure, where an inflection point marks the subspace dimension for which Bayes error ceases to be dominant. To achieve optimality, in the MPE sense, a *retrieval system must therefore operate on the inflection point with the smallest probability of error*.

5.2.2 Optimality criteria

It is straightforward to show (see (52) in the proof of Theorem 1) that a retrieval system with class densities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, and decision function (19) has probability of error

$$P[g(\mathbf{X}) \neq Y] = 1 - E_{\mathbf{X}}[P_{Y|\mathbf{X}}(Y = \arg\max_j \hat{p}_{Y|\mathbf{X}}(j|\mathbf{x})|\mathbf{x})]. \tag{38}$$

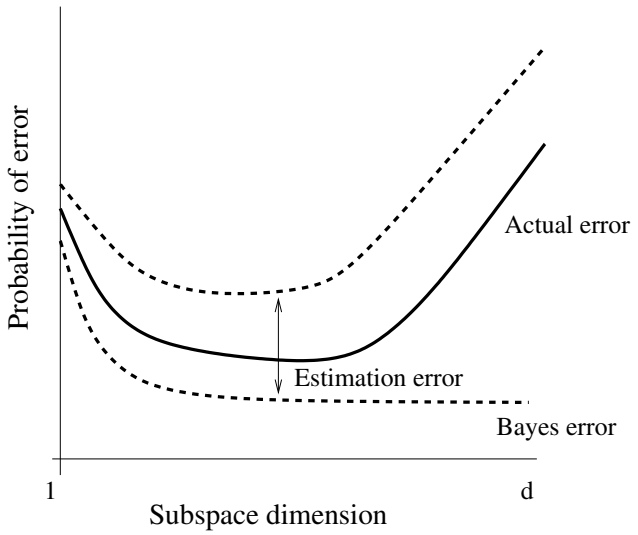

Figure 6: Upper bound, lower bound, and probability of error as a function of subspace dimension.

Nevertheless, because this equation depends on the unknown $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, it is impossible to minimize the probability of error explicitly. One solution is to assume that the estimates $\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)$ are good approximations to the true densities, in which case

$$P[g(\mathbf{X}) \neq Y] \approx 1 - E_{\mathbf{X}}[\max_i \hat{p}_{Y|\mathbf{X}}(i|\mathbf{x})]. \tag{39}$$

In this regime, it follows from the law of large numbers that, given a training sample of image observations $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, the optimal feature transformation is

$$T = \arg\max_A \frac{1}{N} \sum_{k=1}^{N} \max_i \hat{p}_{Y|\mathbf{X}}[i|A(\mathbf{z}_k)]. \tag{40}$$

6 Embedded feature representations

In general, (40) can be a complicated problem since the optimal feature transformation depends on the density estimates $P_{Y|\mathbf{X}}[i|A(\mathbf{z}_k)]$ which, in turn depend on the feature space $\mathcal{X}$. Hence, the optimization must resort to an iterative procedure where densities are estimated for a given feature space and the feature transformation is then updated according to the new estimates. Each of these steps involves cycling through all the image classes in the database and performing operations (e.g. density estimation) which may themselves be non-trivial from a computational standpoint. Since (40) must be solved for each subspace dimension, the procedure is too expensive for most applications.

6.1 Embedded mixture models

A simpler alternative is to consider only sequences of embedded subspaces of a common transformation. The next lemma shows that this can significantly reduce the complexity of density estimation.

Lemma 2 *Let $\mathcal{X}$ be a feature space, $\{\mathcal{X}_j\}$ a sequence of embedded subspaces according to (33), and $\mathbf{X}_1^d$ a sequence of random vectors according to (35). If, under class i, $\mathbf{X}$ is distributed according to the Gaussian mixture density*

$$P_{\mathbf{X}|Y}(\mathbf{x}|i) = \sum_{c=1}^{C} \pi_{i,c}\mathcal{G}(\mathbf{x}, \mu_{i,c}, \mathbf{\Sigma}_{i,c}) \tag{41}$$

then, $\forall j \in 1, \ldots, d$,

$$P_{\mathbf{X}_j|Y}(\mathbf{x}|i) = \sum_{c=1}^{C} \pi_{i,c}\mathcal{G}[\mathbf{x}, \mathbf{\Pi}_j^d\mu_{i,c}, \mathbf{\Pi}_j^d\mathbf{\Sigma}_{i,c}(\mathbf{\Pi}_j^d)^T], \tag{42}$$

where $\mathbf{\Pi}_j^d = [\mathbf{I}_j\ \mathbf{0}_{d-j}]$, is the projection matrix associated with π_j^d, $\mathbf{I}_j$ the $j \times j$ identity matrix, and $\mathbf{0}_{d-j}$ a matrix of zeros.

Proof: see appendix A.7.

The lemma shows that once a set of parameter estimates is obtained for $\mathcal{X}$, the sequence of density estimates in the embedded subspaces $\mathcal{X}_j$ is automatically known. The collection of densities in (42) is denoted by the family of *embedded mixture models* associated with $\mathbf{X}$. Notice that once an estimate is available for $\{\pi_{i,c}, \mu_{i,c}, \mathbf{\Sigma}_{i,c}\}$ the parameters of $P_{\mathbf{X}_j|Y}(\mathbf{x}|i)$ are obtained by simply extracting the first j components of the mean vectors $\mu_{i,c}$ and the upper-left $j \times j$ sub-matrix of the covariances $\mathbf{\Sigma}_{i,c}$. Hence, it is not necessary to repeat the density estimation for each subspace dimension and the overall complexity is really just that of finding the optimal feature transform in $\mathcal{X}$.

In fact, the lemma suggests an efficient cross-validation procedure to find the optimal subspace dimension of a given transformation T. The basic idea is to select a set of query images $I = \{I_1, \ldots, I_Q\}$, establish the associated retrieval ground truth, and use this set to infer the optimal subspace dimension. An algorithmic description of this procedure is given in Figure 7. It remains to determine how the feature transformation T can itself be found. One possibility, that we explore next, is to restrict the search to a finite dictionary of transformations that satisfy some properties known to be important for visual recognition, e.g. invariance to certain image mappings or plausibility under what is known about human perception.

6.2 Embedded multi-resolution mixture models

Ever since the work of Hubel and Wiesel [29], it has been established that 1) visual processing is local, and 2) different groups in primary visual cortex (i.e. area V1) are tuned for detecting different types of stimulus (e.g. bars, edges, and so on). This indicates that, at the lowest level, the architecture of the human visual system can be well

subspacedim$(I, T, \{P_{\mathbf{X}|Y}(\mathbf{x}|i), i = 1, \ldots, M\})$

- *for each query image* $I_s \in I$*:*
  - *apply the transformation* T *to a collection of observations from* I_s *to obtain a set of query feature vectors* $\mathbf{x}_s = \{\mathbf{x}_{s,1}, \ldots, \mathbf{x}_{s,N}\}$
  - *for each subspace dimension* $j = 1, \ldots, d$
    - *for each image class* $i = 1, \ldots, M$
      - *apply (42) to obtain the embedded mixtures* $P_{\mathbf{X}_j|Y}(\mathbf{x}|i)$
      - *compute*
$$p^i_{s,j} = \frac{1}{N} \sum_{k=1}^{N} \log P_{\mathbf{X}_j|Y}(\pi^d_j(\mathbf{x}_{s,k})|i).$$
    - *sort the* $p^i_{s,j}$ *by decreasing value and, based on the resulting order, evaluate some measure of retrieval performance (e.g. precision at some level of recall)* $R_{s,j}$.
- *average the retrieval measure across queries* $R_j = 1/|I| \sum_s R_{s,j}$.
- *return the subspace dimension* $j^* = \arg\max_j R_j$ *and associated performance score* R_{j^*}.

Figure 7: Algorithm for determining the optimal subspace dimension for a retrieval problem with feature transformation T, and class densities $\{P_{\mathbf{X}|Y}(\mathbf{x}|i), i = 1, \ldots, M\}$.

approximated by a multi-resolution representation localized in space and frequency, and several "biologically plausible" models of early vision are based on this principle [60, 42, 4, 21, 67, 5].

A space/space-frequency representation is obtained by convolving the image with a collection of elementary filters of reduced spatial support and tuned to different spatial frequencies and orientations. Several elementary filters have been proposed in the literature, including *differences of Gaussians* [42], *Gabor functions* [55, 21], and *differences of offset Gaussians* [42], among others. More recently, it has been shown that filters remarkably similar to the receptive fields of cells found in V1 [50, 2] can be learned from training images, by imposing requirements of sparseness [20, 50] or independence [2] in the space/space-frequency coefficients.

When the feature transform T is a multi-resolution decomposition embedded mixture densities have an interesting interpretation as families of densities defined over multiple image scales, each adding higher resolution information to the characterization provided by those before it. In fact, disregarding the dimensions associated with high-frequency basis functions is equivalent to modeling densities of low-pass filtered images. In the extreme case where only the first, or DC, coefficient is considered the representation is equivalent to the histogram of a smoothed version of the original image. This is illustrated in Figure 8.

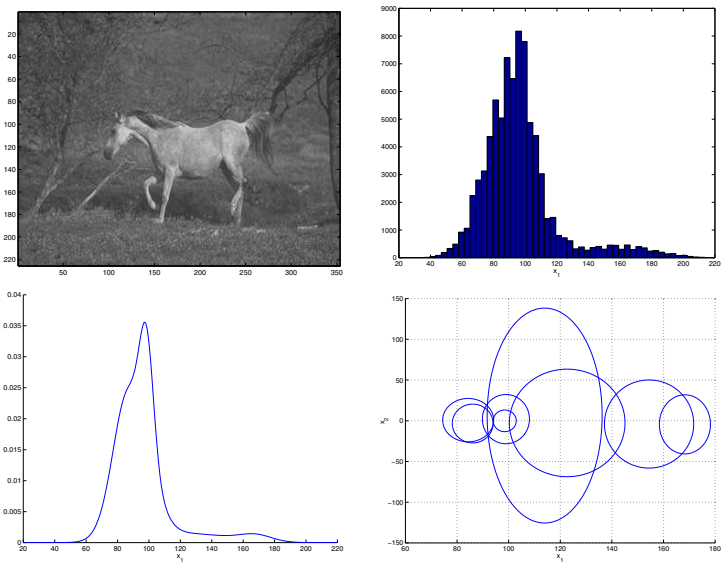

Figure 8: An image from the Corel database (top left), its histogram (top right), and projections of the corresponding 64-dimensional embedded mixture onto the DC subspace (bottom left), and the subspace of the two lower frequency coefficients (bottom right).

The *embedded multi-resolution mixture* (EMM) model (embedded mixtures on a multi-resolution feature space) can thus be seen as a *generalization of the color histogram, where the additional dimensions capture the spatial dependencies that are crucial for fine image discrimination* (as illustrated in Figure 4). This generalization also enables fine control over the invariance properties of the representation. Since the histogram is approximately invariant to scaling, rotation, and translation, when only the DC subspace is considered the representation is invariant to all these transformations. As high-frequency coefficients are included, invariance is gradually sacrificed. Of course, invariance can always be improved by including the proper examples in the training sample used to learn the parameters of the model.

6.3 Optimal features

Given a finite collection $\mathcal{T} = \{T^{(1)}, \ldots, T^{(F)}\}$ of multi-resolution transformations, the optimal transformation can be found by exhaustive search based on the algorithm of Figure 7. In this case, the only non-trivial issue is how to efficiently estimate the densities $\{P_{\mathbf{X}|Y}(\mathbf{x}|i), i = 1, \ldots, M\}$ on the different feature spaces. Notice that if $T^{(l)} : \mathcal{Z} \to \mathcal{X}^{(l)}$ and $T^{(m)} : \mathcal{Z} \to \mathcal{X}^{(m)}$ are two invertible transformations in $\mathcal{T}$, then the transformation $T^{(l,m)} = T^{(m)} \circ (T^{(l)})^{-1}$ maps $\mathcal{X}^{(l)}$ into $\mathcal{X}^{(m)}$. It follows, using arguments similar to those of the proof of Lemma 2, that if in $\mathcal{X}^{(l)}$ the feature

optimal transform$(I, \mathcal{T})$

1. *select a reference transformation in $\mathcal{T}$, e.g. $T^{(1)}$;*
2. *for each image class $i = 1, \ldots, M$, use a standard maximum likelihood estimation technique, e.g. the expectation-maximization algorithm [14], to determine the mixture parameters of $P_{\mathbf{X}^{(1)}|Y}(\mathbf{x}|i)$;*
3. *for each transformation $f = 2, \ldots, F$*
   - *let $T^{(1,f)} = T^{(f)} \circ (T^{(1)})^{-1}$*
   - *compute, for each image class $i = 1, \ldots, M$, the parameters of $P_{\mathbf{X}^{(f)}|Y}(\mathbf{x}|i)$ using (45) and (46).*
   - *let $(j_f^*, R_f^*) =$* **subspacedim**$(I, T^{(f)}, \{P_{\mathbf{X}^{(f)}|Y}(\mathbf{x}|i), i = 1, \ldots, M\})$
4. *let $f^* = \arg\max_f R_f^*$ and $j^* = j_{f^*}^*$;*
5. *return $T_{j^*}^{(f^*)} = \pi_{j^*}^d(T^{(f^*)})$*

Figure 9: Algorithm for determining the best feature transformation, and subspace dimension for a retrieval problem with transformation dictionary $\mathcal{T}$.

distribution is, for class i,

$$P_{\mathbf{X}^{(l)}|Y}(\mathbf{x}|i) = \sum_{c=1}^{C} \pi_{i,c} \mathcal{G}(\mathbf{x}, \mu_{i,c}^l, \mathbf{\Sigma}_{i,c}^l) \tag{43}$$

then, on $\mathcal{X}^{(m)}$,

$$P_{\mathbf{X}^{(m)}|Y}(\mathbf{x}|i) = \sum_{c=1}^{C} \pi_{i,c} \mathcal{G}(\mathbf{x}, \mu_{i,c}^m, \mathbf{\Sigma}_{i,c}^m) \tag{44}$$

where

$$\mu_{i,c}^m = \mathbf{T}^{(l,m)} \mu_{i,c}^l \tag{45}$$

$$\mathbf{\Sigma}_{i,c}^m = \mathbf{T}^{(l,m)} \mathbf{\Sigma}_{i,c}^l (\mathbf{T}^{(l,m)})^T . \tag{46}$$

Therefore, it suffices to perform density estimation on a reference subspace, e.g. $\mathcal{X}^{(1)}$, in order to obtain the mixture parameters associated with all transformations in $\mathcal{T}$. The search for the optimal feature transformation can thus be performed with the algorithm of Figure 9.

6.4 Multi-resolution feature transforms

For a feature transformation $T : \mathcal{Z} \to \mathcal{X}$ one can define an inverse, reconstruction, mapping

$$A : \mathcal{X} \to \mathcal{Z}.$$

The columns of the associated matrix $\mathbf{A}$ are called basis functions of the transformation. When $\mathbf{A} = \mathbf{T}^T$ the transformation is orthogonal. Various popular space/space-frequency representations are derived from orthogonal feature transforms.

Definition 6 *The Discrete Cosine Transform (DCT) [32] of size n is the orthogonal transform whose basis functions are defined by:*

$$A(i,j) = \alpha(i)\alpha(j)\cos\frac{(2x+1)i\pi}{2n}\cos\frac{(2y+1)j\pi}{2n},\ 0 \le i,j,x,y < n \tag{47}$$

where $\alpha = \sqrt{1/n}$ for $i = 0$, and $\alpha = \sqrt{2/n}$ otherwise.

The DCT is widely used in image compression, and previous recognition experiments have shown that DCT features can lead to recognition rates comparable to those of many features proposed in the recognition literature [77]. It is also possible to show that, for certain classes of stochastic processes, the DCT converges asymptotically to the following transform [32].

Definition 7 *Principal Components Analysis (PCA) is the orthogonal transform defined by*

$$\mathbf{T} = \mathbf{D}^{-1/2}\mathbf{E}^T, \tag{48}$$

where $\mathbf{EDE}^T$ is the eigenvector decomposition of the covariance matrix $E[\mathbf{z}\mathbf{z}^T]$.

It is well known (and straightforward to show) that PCA generates uncorrelated features, i.e. $E[\mathbf{x}\mathbf{x}^T] = \mathbf{I}$. In this context, PCA is the optimal redundancy reduction transform, i.e. the one that produces the most parsimonious description of the input observations. For this reason, PCA has been widely used in both compression and recognition [74, 48].

While they originate spatial/spatial-frequency representations, the major limitation of the above transforms as models for visual perception is the arbitrary nature of their spatial localization (enforced by arbitrarily segmenting images into blocks). This can result in severe scaling mismatches if the block size does not match that of the image detail. Such scaling problems are alleviated by the wavelet representation.

Definition 8 *A wavelet transform (WT) [43] is the orthogonal transform whose basis functions are defined by*

$$A(i,j) = \sqrt{2^{k+l}}\Psi\left(2^k x - i\right)\Psi\left(2^l y - j\right)\ {}^{0\le k,l<\log_2 n}_{(0,0)\le(i,j)<(2^k,2^l)} \tag{49}$$

where $\Psi(x)$ is a function (wavelet) that integrates to zero.

Like the DCT, wavelets have been shown empirically to achieve good decorrelation. However, natural images exhibit a significant amount of higher-order dependencies that cannot be captured by orthogonal components [50]. Eliminating such dependencies is the goal of independent component analysis.

Definition 9 *Independent Component Analysis (ICA) [10] is a feature transform such that*

$$P_{\mathbf{X}}(\mathbf{x}) = \prod_i P_{\mathbf{X}_i}(\mathbf{x}_i) \tag{50}$$

where $\mathbf{X} = (X_1, \ldots, X_d)$ *is the random process from which feature vectors are drawn.*

The exact details of ICA depend on the particular algorithm used to learn the basis from a training sample. Since independence is usually difficult to measure and enforce if d is large, ICA techniques tend to settle for less ambitious goals. The most popular solution is to minimize a contrast function which is guaranteed to be zero if the inputs are independent. Examples of such contrast functions are higher order correlations and information-theoretic objective functions[10]. In this work, we consider representatives from the two types: the method developed by Comon [12], which uses a contrast function based on high-order cumulants, and the FastICA algorithm [30], that relies on the negative entropy of the features.

7 Experimental evaluation

In this section, we present an experimental evaluation of DTR and a comparison against various retrieval techniques in common use. In the retrieval context, it is desirable to rely on a generic representation that can achieve equally good performance for diverse types of imagery. For this reason, we conducted experiments on three different databases: the Brodatz texture database, the Columbia object database, and a subset of the Corel database of stock photography. While Brodatz provides a good testing ground for texture retrieval, color-based methods tend to do well on Columbia. Corel contains generic imagery and requires retrieval algorithms that can account for both color and texture. In each case, we surveyed the literature to identify a competing technique that is representative of the state of the art in each area.

7.1 Databases and performance evaluation

The 1008 images in the Brodatz database were divided into two subgroups: a *query database* of 112, and a *retrieval database* of 896 images. Various previous studies have identified the combination of 1) the coefficients of the least squares fit of a *multi-resolution simultaneous auto-regressive* (MRSAR) model to each texture and 2) the Mahalanobis distance, as a top performer in this database [54, 40, 41]. Following [45, 40, 44], the MRSAR features were computed using a window of size 21×21 sliding over the image with increments of two pixels in both the horizontal and vertical dimensions. Each feature vector consists of 4 SAR parameters plus the error of the fit achieved by the SAR model at three resolutions, in a total of 15 dimensions.

The Columbia database was also split into two subsets: a query database containing a single view of each of the 100 objects available, and a retrieval database containing 9 views (separated by 40^o) of each object. It was chosen because it is a database where histogram-based methods tend to perform well, allowing a comparison of DTR against these techniques. For color histogramming, the 3D color space was quantized

by finding the bounding box for all the points in the query and retrieval databases and then dividing each axis in b bins. This leads to b^3 cells. Experiments were performed with different values of b. Retrieval was based on HI.

From Corel we selected 15 image classes[1] leading to a total of $1,500$ images. Of these, 20% were used on the query database, leaving the remaining 80% for retrieval. In addition to the texture and color-based approaches, retrieval performance was compared against those of two other approaches that are representative of the state of the art in the joint modeling of the two attributes: the color correlogram [28] and the linear combination of color and texture distances. To combine color and texture distances linearly we started by evaluating all the distances between query and database entries according to both HI and MRSAR/MD. For each query, we then normalized all distances by their mean and variance, clipped all values with magnitude larger than three standard deviations, and mapped the resulting interval into $[0, 1]$. An overall distance was then computed for each entry in the database, according to

$$d' = (1 - w) \times d_c + w \times d_t$$

where d_c and d_t are, respectively, the normalized distances according to HI and MRSAR/MD, and w a pre-defined weight.

On Corel and Columbia images were converted from the original RGB to the YBR color space. The feature space had 64 dimensions per color channel, and features from the different channels were interleaved according to the pattern YBRYBR... Mixtures of 8 Gaussians were used for the Brodatz and Corel databases and 16 for Columbia. Diagonal covariances were used for all Gaussians, and all the mixture parameters were learned with the EM algorithm [14]. Each image in the database was considered as an independent class. A series of experiments were designed with the goal of evaluating the performance of the individual components of the DTR architecture.

7.2 Similarity function

The first series of experiments was designed to compare the performance of the Bayesian retrieval criteria with that of the, more commonly used, MD or HI similarity functions. In order to isolate the contribution of the similarity function from those of the features and the feature representation, the comparison was performed with the feature sets and representations discussed above for the Brodatz and Columbia databases: color-based retrieval was implemented by combining the color histogram with (8) and texture-based retrieval by the combination of the MRSAR features with (15).

Figure 10 presents precision/recall (PR) curves for the two databases. As expected, texture-based retrieval (MRSAR) performs better on Brodatz while color-based retrieval (color histogramming) does better on Columbia. However, when the appropriate features and representation are used for the specific database, ML always leads to a clear improvement in retrieval precision. In particular, for the texture database, combining ML with the MRSAR features and the Gaussian representation leads to an

[1] "Arabian horses", "auto racing", "coasts", "divers and diving", "English country gardens", "fireworks", "glaciers and mountains", "Mayan and Aztec ruins", "oil paintings", "owls", "land of the pyramids", "roses", "ski scenes", "religious stained glass".

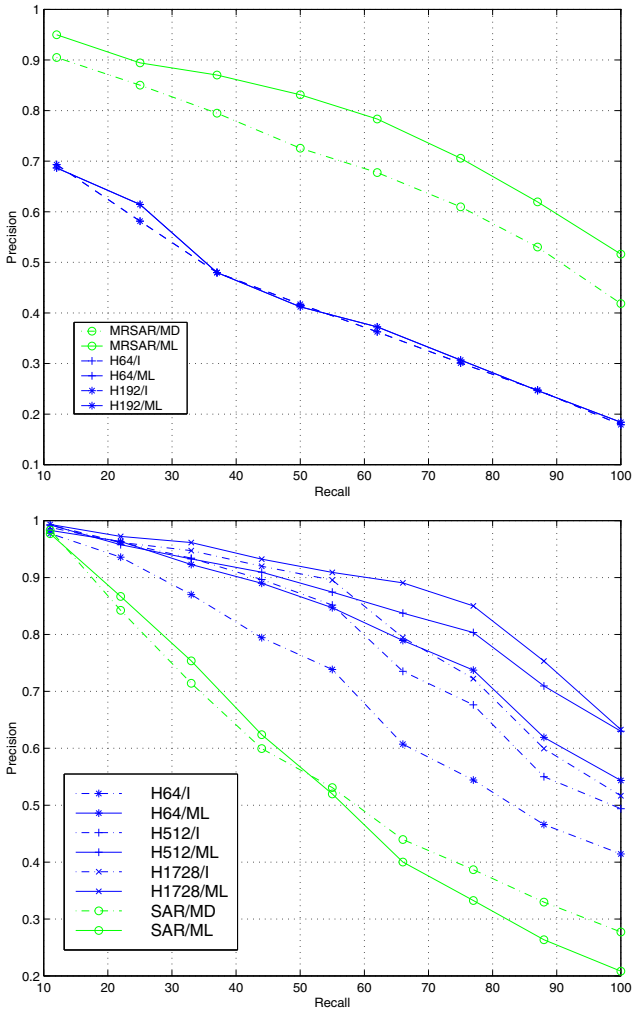

Figure 10: PR curves for Brodatz (top) and Columbia (bottom). In the legend, MRSAR means MRSAR features, H color histograms, ML maximum likelihood, MD Mahalanobis distance, and I intersection. The total number of bins in each histogram is indicated after the H.

improvement in precision from 5 to 10% (depending on the level of recall) over that achievable with the MD. Similarly, on Columbia, replacing HI by the ML criteria leads to an improvement that can be as high as 20%.

7.3 Feature transformation

The next set of experiments was designed to assess the retrieval performance of the various multi-resolution transformations discussed in section 6.4. For this, we measured precision at various levels of recall, on Brodatz and Corel. Since Corel is a larger database and contains colored images, 192-dimensional feature space, the queries take longer to compute. For this reason, we restricted the analysis to the first 64 dimensions (and only considered one of the ICA techniques). In both databases, the relative precision of the various transformations was similar at all levels of recall. We, therefore, only present curves of precision, as a function of subspace dimension, at 30% recall on Brodatz and 10% recall on Corel. These are shown in Figure 11.

Notice that the precision curves comply with the theoretical arguments of section 5.2.1. Since precision is inversely proportional to the probability of error one would expect, from those arguments, the precision curves to be concave. This is indeed the case (there is a large increase in precision from 1 to 8 dimensions that we do not show for clarity of the graph) for all transformations.

In terms of the relative performance of the different transforms, the DCT is the top performer for both databases reaching high precision in both cases. On the other hand, PCA always performs poorly. This is a significant result, given that PCA has been widely used in visual recognition [74, 48]. The performance of the other features seems to be significantly more dependent on the database. Wavelets do quite well on Corel, but very poorly on Brodatz, ICA does better on Brodatz than on Corel.

It is important to emphasize that, for a given database, a poor choice of transformation can lead to significant degradation of retrieval performance. On Brodatz the peak precision of the worst transformation (wavelet) is 10% below that of the best (DCT), on Corel the variation is almost 20%. Even for a given transformation, precision can vary dramatically with the number of embedded subspaces. For example, the precision of the DCT features on Brodatz drops from the peak value of about 92% to about 62% when all the subspaces are included. These observations emphasize the importance of the feature selection algorithm discussed in section 6.3.

7.4 Comparison to standard solutions

Since the DCT features achieved the best performance across databases, they were the only features considered in the remaining experiments. In this section, we present a comparison of the performance of EMM/ML/DCT combination against those of MRSAR/ML and HI, in the specific databases where the latter work best: texture (Brodatz) for MRSAR and color (Columbia) for HI. Figure 12 presents the resulting PR curves, showing that EMM/ML/DCT achieves equivalent performance or actually outperforms the best of the two other approaches in each image domain. This demonstrates that, despite its generic nature, the EMM/ML/DCT representation can handle both color and texture and should therefore do well on a large spectrum of databases.

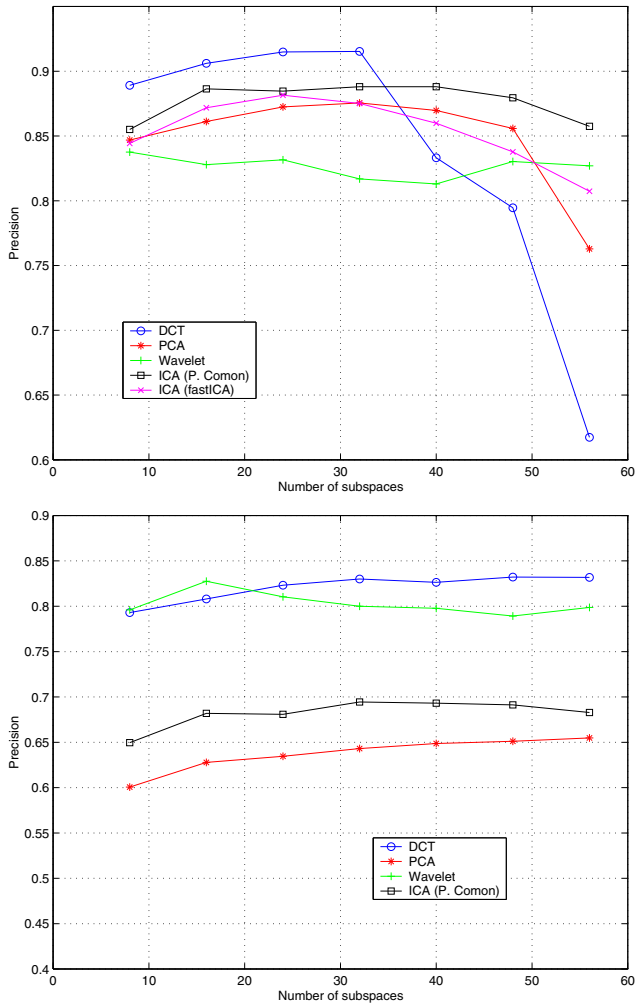

Figure 11: Top: Precision, at 30% recall, on Brodatz. Bottom: Precision, at 30% recall, on Corel.

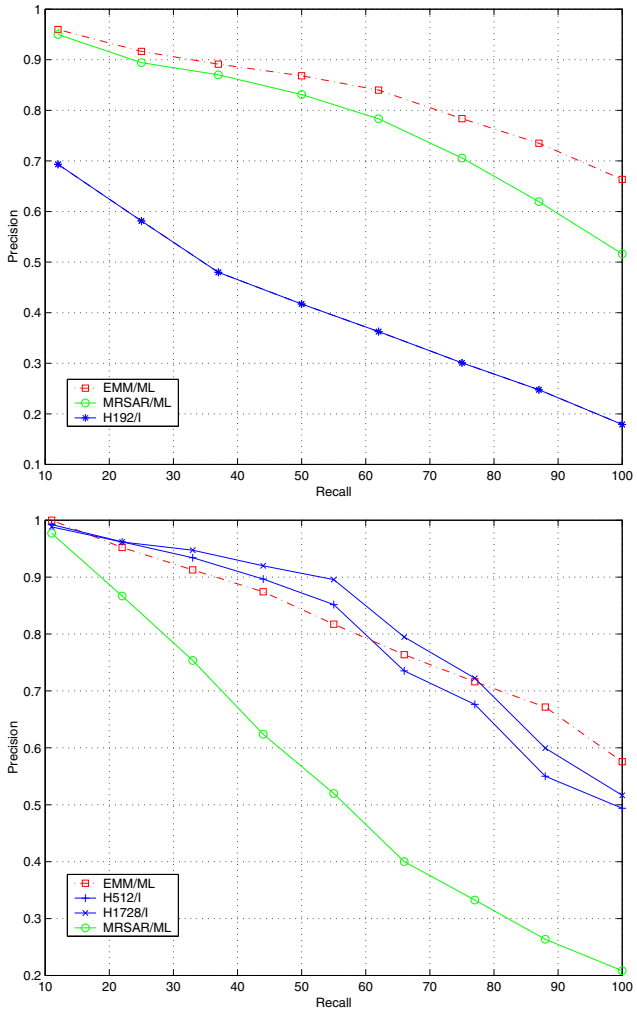

Figure 12: PR curves of EMM/ML/DCT, MRSAR/ML, and HI on Brodatz (top) and Columbia (bottom).

Visual inspection of the retrieval results suggests that, also along the dimension of perceptual relevance, EMM/ML/DCT clearly beats the MRSAR and histogram-based approaches. Figure 13, presents representative examples of the three major advantages of EMM/ML[2]: 1) when it makes errors, these tend to be perceptually less annoying than those of the other approaches, 2) when there are several visually similar classes in the database, images from these classes tend to be retrieved together, and 3) even when the performance is worse than that of the other approaches in terms of PR, the results are frequently better from a perceptual standpoint.

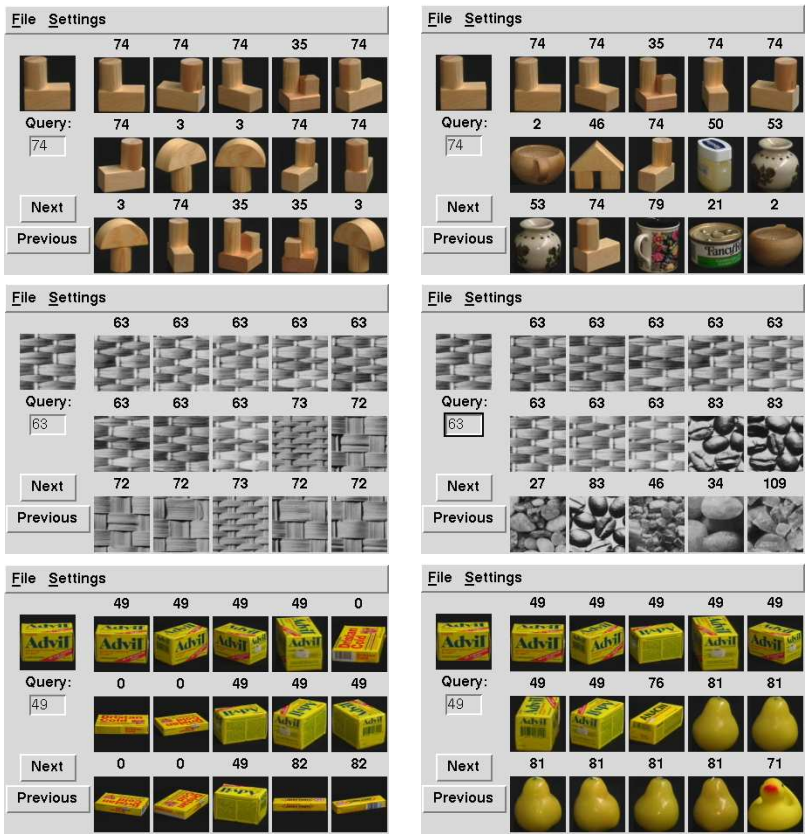

Figure 13: Comparison of EMM/ML/DCT retrieval results (left column) with those of HI on Columbia and MRSAR/MD on Brodatz (right column). The number above each image indicates the class to which it belongs.

The two pictures on the top row exemplify how EMM/ML/DCT can lead to perceptually pleasing retrieval results, even when the PR performance is only mediocre.

[2]A significantly larger set of retrieval examples is available from `http://crl.research.compaq.com/vision/multimedia/retrieval/default.htm`, and a more detailed analysis presented in [77].

In this case, while HI retrieves several objects unrelated to the query, EMM/ML/DCT only returns objects that, like the query, are made of wood blocks. This is due to the fact that, by relying on features with spatial support, EMM/ML/DCT is able to capture the local appearance of the object surface. It will thus tend to match surfaces with the same shape, texture, and reflection properties. This is not possible with color histograms.

The two images on the center exemplify situations where both approaches perform perfectly in terms of PR, yet the perceptual retrieval quality is very different. MRSAR/ML ranks all the images in the query class at the top, but produces non-sense matches after that. On the other hand, EMM/ML/DCT retrieves images that are visually similar to the query after all the images in its class are exhausted. This observation is frequent and derives from the fact that the MRSAR features have no perceptual justification.

Finally, the pictures on the bottom illustrate how, even when it has higher PR, HI frequently leads to perceptually poorer results than EMM/ML/DCT. In this case, images of a pear and a duck are retrieved by HI after the images in the right class ("Advil box"), even though there are several boxes with colors similar to those of the query in the database. On the other hand, EMM/ML/DCT only retrieves boxes, although not in the best possible order.

7.5 Generic retrieval

We finalize with results from Corel. The top plot on Figure 14 presents a comparison, in terms of PR, of MRSAR/MD, HI, the color correlogram, and EMM/ML/DCT. It is clear that the texture model alone performs very poorly, color histogramming does significantly better, and the correlogram further improves performance by about 5%. However, all these approaches are significantly less effective than EMM/ML/DCT. Similarly, the bottom plot compares the PR curves of EMM/ML/DCT with those obtained by linear weighting of the color and texture distances. Several curves are shown for values of $w \in [0, 1]$. It is clear that linear weighting is never better than EMM/ML/DCT. Given that, in a realistic retrieval scenario, the value of the optimal weight is not known, there are no simple ways to determine it, and the linear combination always requires an increase in complexity (distances have to be computed according to the two representations), we see no reason to prefer this type of solution to DTR.

We conclude with three retrieval examples, shown in Figure 15, from the Corel database. These examples illustrate the robustness of DTR to changes in scene layout, object position and orientation, or variability in the background. A significantly larger set of retrieval examples is available from
`http://crl.research.compaq.com/vision/multimedia/retrieval/default.htm`.

8 Acknowledgements

I would like to thank various people that have contributed to this work. Andy Lippman for first challenging me to solve the retrieval problem and providing lots of initial guidance; Bob Gray for invaluable suggestions and feedback on the theoretical aspects of the work; Murat Kunt, Roz Picard, and Aaron Bobick and the members of the

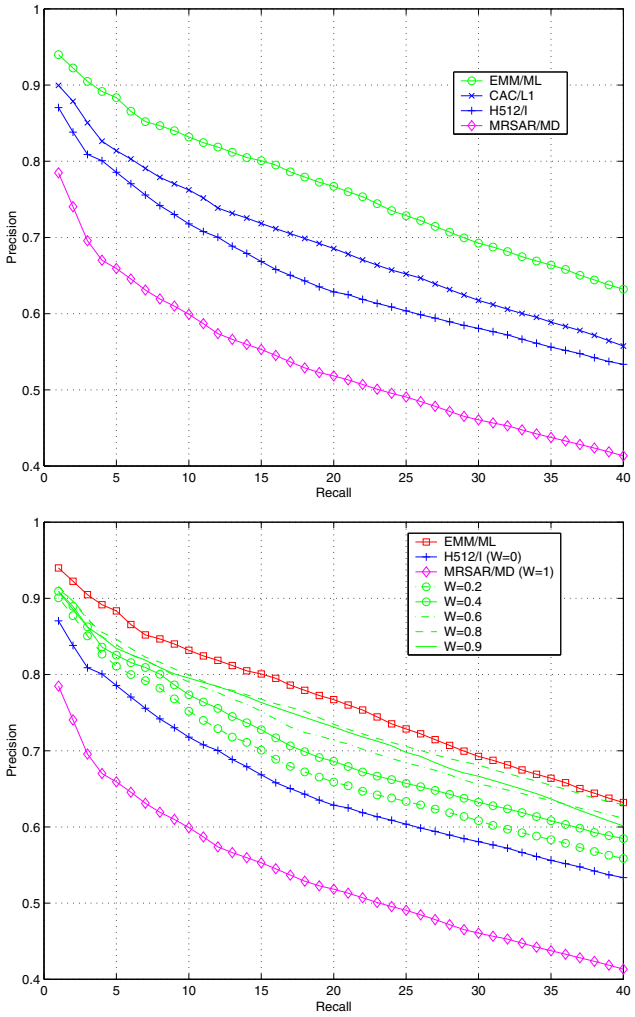

Figure 14: Top: PR on Corel for MRSAR/MD, HI, color correlogram (CAC), and EMM/ML/DCT. Bottom: Comparison of PR achieved with EMM/ML/DCT and linear weighting of MRSAR/MD and HI for different weights.

Figure 15: Queries for “Arabian horses”, “auto racing”, and “English gardens”.

VIP, vision, and speech groups at CRL for various interesting discussions; and Gustavo Carneiro for an extremely productive summer where we performed many of the experiments reported in the paper.

A Proofs of the theoretical results

For any two integers i and j, the *Kronecker delta* function is defined by

$$\delta_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{51}$$

A.1 Proof of theorem 1

Proof: The proof can be found in various textbooks (see [15, 22] among many others). We include it here because some of the intermediate results are required by subsequent proofs.

The probability of error associated with the decision rule $g(\mathbf{x})$ is

$$\begin{aligned} P_{\mathbf{X},Y}(g(\mathbf{X}) \neq Y) &= \int P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= E_{\mathbf{x}}[P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x})], \end{aligned} \tag{52}$$

where

$$\begin{aligned} P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) &= \sum_i P(Y \neq g(\mathbf{X})|\mathbf{X} = \mathbf{x}, Y = i) P_{Y|\mathbf{X}}(i|\mathbf{x}) \\ &= \sum_i (1 - \delta_{g(\mathbf{x}),i}) P_{Y|\mathbf{X}}(i|\mathbf{x}) \\ &= 1 - \sum_i \delta_{g(\mathbf{x}),i} P_{Y|\mathbf{X}}(i|\mathbf{x}) \end{aligned} \tag{53}$$

and $\delta_{i,j}$ is the Kronecker delta function defined in (51). It follows that

$$\begin{aligned} P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) &\geq 1 - \max_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \\ &= 1 - P_{Y|\mathbf{X}}(Y = g^*(\mathbf{x})|\mathbf{x}) \\ &= P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x}) \end{aligned}$$

and, consequently

$$E_{\mathbf{x}}[P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x})] \geq E_{\mathbf{x}}[P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x})].$$

I.e., any other decision rule will have a larger probability of error than the Bayes classifier. Since, from (52),

$$\begin{aligned} P_{\mathbf{X},Y}(g^*(\mathbf{X}) \neq Y) &= 1 - E_{\mathbf{x}}[P_{Y|\mathbf{X}}(Y = g^*(\mathbf{x})|\mathbf{x})] \\ &= 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})] = L^* \end{aligned}$$

the probability of error can never be smaller than the Bayes error.

A.2 Proof of theorem 2

Proof: The following proof is a straightforward extension to multiple classes of the one given in [15] for the two-class problem. From (5),

$$\begin{aligned} L^*_{\mathcal{X}} &= 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \\ &= 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|T(\mathbf{z}))], \\ &= 1 - E_{\mathbf{x}}[\max_i \int P_{Y|\mathbf{Z},\mathbf{X}}(i|\mathbf{z}, T(\mathbf{z}))P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|T(\mathbf{z}))d\mathbf{z}], \\ &= 1 - E_{\mathbf{x}}[\max_i \int P_{Y|\mathbf{Z}}(i|\mathbf{z})P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|T(\mathbf{z}))d\mathbf{z}], \\ &= 1 - E_{\mathbf{x}}[\max_i E_{\mathbf{z}|\mathbf{X}}[P_{Y|\mathbf{Z}}(i|\mathbf{z})|\mathbf{X} = T(\mathbf{z})]], \\ &\geq 1 - E_{\mathbf{x}}[E_{\mathbf{z}|\mathbf{X}}[\max_i P_{Y|\mathbf{Z}}(i|\mathbf{z})|\mathbf{X} = T(\mathbf{z})]], \\ &= 1 - E_{\mathbf{z}}[\max_i P_{Y|\mathbf{Z}}(i|\mathbf{z})] = L^*_{\mathcal{Z}}, \end{aligned}$$

where equality is achieved if and only if T is an invertible map, and we have used the fact that

$$E_{\mathbf{z}|\mathbf{X}}[\max_i P_{Y|\mathbf{Z}}(i|\mathbf{z})|\mathbf{X} = T(\mathbf{z})] \geq E_{\mathbf{z}|\mathbf{X}}[P_{Y|\mathbf{Z}}(i|\mathbf{z})|\mathbf{X} = T(\mathbf{z})], \forall i.$$

A.3 Proof of theorem 3

Proof: From (52),

$$P_{\mathbf{X},Y}(g(\mathbf{X}) \neq Y) - L^*_{\mathcal{X}} =$$

$$\int [P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) - P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x})]P_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \tag{54}$$

and since, $\forall \mathbf{x} \in \mathcal{X}$ such that $g(\mathbf{x}) = g^*(\mathbf{x})$, we have

$$P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) = P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x}),$$

this is equivalent to

$$P_{\mathbf{X},Y}(g(\mathbf{X}) \neq Y) - L^*_{\mathcal{X}} =$$

$$\int_E [P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) - P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x})] P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x},$$

where

$$E = \{\mathbf{x}|\mathbf{x} \in \mathcal{X}, P_{\mathbf{X}}(\mathbf{x}) > 0, g(\mathbf{x}) \neq g^*(\mathbf{x})\}.$$

Letting

$$\Delta(\mathbf{x}) = P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) - P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x})$$

and defining the sets

$$E^*_i = \{\mathbf{x}|\mathbf{x} \in E, g^*(\mathbf{x}) = i\}$$

$$E_i = \{\mathbf{x}|\mathbf{x} \in E, g(\mathbf{x}) = i\},$$

it follows from (53) that, $\forall \mathbf{x} \in E^*_i \cap E_j$,

$$\Delta(\mathbf{x}) = P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}).$$

Since, from (4),

$$P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}) \geq 0 \ \forall \mathbf{x} \in E^*_i, \forall j \neq i,$$

from (19) and the fact that $P_{\mathbf{X}}(\mathbf{x}) > 0 \, \forall \mathbf{x} \in E$,

$$\frac{\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|j)\hat{p}_Y(j)}{P_{\mathbf{X}}(\mathbf{x})} - \frac{\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i)}{P_{\mathbf{X}}(\mathbf{x})} \geq 0 \ \forall \mathbf{x} \in E_j, \forall i \neq j,$$

defining

$$\hat{p}_{Y|\mathbf{X}}(i|\mathbf{x}) = \frac{\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i)}{P_{\mathbf{X}}(\mathbf{x})},$$

we have, $\forall \mathbf{x} \in E^*_i \cap E_j$,

$$\begin{aligned}
\Delta(\mathbf{x}) &= P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}) \\
&\leq P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}) + \hat{p}_{Y|\mathbf{X}}(j|\mathbf{x}) - \hat{p}_{Y|\mathbf{X}}(i|\mathbf{x}) \\
&= |P_{Y|\mathbf{X}}(i|\mathbf{x}) - P_{Y|\mathbf{X}}(j|\mathbf{x}) + \hat{p}_{Y|\mathbf{X}}(j|\mathbf{x}) - \hat{p}_{Y|\mathbf{X}}(i|\mathbf{x})| \\
&\leq |P_{Y|\mathbf{X}}(i|\mathbf{x}) - \hat{p}_{Y|\mathbf{X}}(i|\mathbf{x})| + |P_{Y|\mathbf{X}}(j|\mathbf{x}) - \hat{p}_{Y|\mathbf{X}}(j|\mathbf{x})|
\end{aligned}$$

and

$$\begin{aligned}\int_{E_i^* \cap E_j} \Delta(\mathbf{x}) P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} &\leq \int_{E_i^* \cap E_j} |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x} \\ &+ \int_{E_i^* \cap E_j} |P_{\mathbf{X}|Y}(\mathbf{x}|j) P_Y(j) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|j) \hat{p}_Y(j)| d\mathbf{x}.\end{aligned}$$

Using the fact that both collections of sets E_i^* and E_j partition E,

$$\begin{aligned}\int_E \Delta(\mathbf{x}) P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} &= \sum_{i,j} \int_{E_i^* \cap E_j} \Delta(\mathbf{x}) P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &\leq \sum_i \int_{E_i^*} |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x} + \\ &\quad \sum_j \int_{E_j} |P_{\mathbf{X}|Y}(\mathbf{x}|j) P_Y(j) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|j) \hat{p}_Y(j)| d\mathbf{x} \\ &= \sum_i \left[\int_{E_i^*} |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x} \right. \\ &\quad \left. + \int_{E_i} |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x} \right] \\ &\leq \sum_i \int |P_{\mathbf{X}|Y}(\mathbf{x}|i) P_Y(i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i) \hat{p}_Y(i)| d\mathbf{x}\end{aligned}$$

where we have also used the fact that $E_i^* \cap E_i = \emptyset$.

A.4 Proof of Corollary 1

Proof: This is a straightforward consequence of applying Pinsker's inequality [13]

$$\int |P_{\mathbf{X}}(\mathbf{x}) - Q_{\mathbf{X}}(\mathbf{x})| d\mathbf{x} \leq KL[P_{\mathbf{X}}(\mathbf{x}) || Q_{\mathbf{X}}(\mathbf{x})] \tag{55}$$

to (20) under the assumption of equiprobable classes.

A.5 Proof of Lemma 1

Proof: From the well known fact that, for any densities P and Q, $KL[P||Q] \geq 0$ [13] it follows that $\Delta_{g,\mathcal{X}}$ is minimum if and only if, for all i

$$\mathbf{p}_i = \arg\min_{\mathbf{p}} KL[P_{\mathbf{X}|Y}(\mathbf{x}|i) || P_{\mathbf{X}|Y}(\mathbf{x}|i; \mathbf{p})] \tag{56}$$

$$= \arg\max_{\mathbf{p}} \int P_{\mathbf{X}|Y}(\mathbf{x}|i) \log P_{\mathbf{X}|Y}(\mathbf{x}|i;\mathbf{p}) d\mathbf{x} \quad (57)$$

$$= \arg\max_{\mathbf{p}} E_{\mathbf{X}_i}[\log P_{\mathbf{X}|Y}(\mathbf{x}|i;\mathbf{p})]. \quad (58)$$

The lemma follows by application of the law of large numbers.

A.6 Proof of Theorem 4

Proof: The fact that the sequence of vector spaces is embedded follows from (33) since, $\forall i \in \{1, \ldots, d-1\}$

$$\mathcal{X}_i = \pi_i^{i+1}(\mathcal{X}_{i+1}) \quad (59)$$

and, consequently,

$$\iota_i^{i+1}(\mathcal{X}_i) \subset \mathcal{X}_{i+1}. \quad (60)$$

Inequality (34) then follows from (59), (18) and the fact that the mappings $\pi_i^{i+1}(\mathbf{x})$ are non-invertible. To prove (37) we start from Corollary 1, i.e.

$$\Delta_{g_i,\mathcal{X}_i} = \sum_k KL[P_{\mathbf{X}_i|Y}(\mathbf{x}|k)||\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}|k)], \quad (61)$$

where $P_{\mathbf{X}_i|Y}(\mathbf{x}|k)$ is the class-conditional likelihood function for $\mathbf{X}_i$ under class k. Since, from (59), $\mathbf{X}_{i+1} = (\mathbf{X}_i, X_{i+1})$ where X_{i+1} is the $i+1^{th}$ coordinate of $\mathbf{X}_{i+1}$

$$KL[P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k)||\hat{p}_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k)] =$$

$$= \int P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k) \log \frac{P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k)}{\hat{p}_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k)} d\mathbf{x}$$

$$= \int P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k) \log \frac{P_{X_{i+1}|\mathbf{X}_i,Y}(x_{i+1}|\pi_i^{i+1}(\mathbf{x}),k)}{\hat{p}_{X_{i+1}|\mathbf{X}_i,Y}(x_{i+1}|\pi_i^{i+1}(\mathbf{x}),k)} d\mathbf{x}$$

$$+ \int P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k) \log \frac{P_{\mathbf{X}_i|Y}(\pi_i^{i+1}(\mathbf{x})|k)}{\hat{p}_{\mathbf{X}_i|Y}(\pi_i^{i+1}(\mathbf{x})|k)} d\mathbf{x}$$

$$= \int P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k) \log \frac{P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k)}{\hat{p}_{X_{i+1}|\mathbf{X}_i,Y}(x_{i+1}|\pi_i^{i+1}(\mathbf{x}),k) P_{\mathbf{X}_i|Y}(\pi_i^{i+1}(\mathbf{x})|k)} d\mathbf{x}$$

$$+ \int P_{\mathbf{X}_i|Y}(\mathbf{x}|k) \log \frac{P_{\mathbf{X}_i|Y}(\mathbf{x}|k)}{\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}|k)} d\mathbf{x}$$

$$= KL[P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}|k)||\hat{p}_{X_{i+1}|\mathbf{X}_i,Y}(x_{i+1}|\pi_i^{i+1}(\mathbf{x}),k) P_{\mathbf{X}_i|Y}(\pi_i^{i+1}(\mathbf{x})|k)]$$

$$+ KL[P_{\mathbf{X}_i|Y}(\mathbf{x}|k)||\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}|k)]$$

$$\geq \quad KL[P_{\mathbf{X}_i|Y}(\mathbf{x}|k)||\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}|k)]$$

where we have used the non-negativity of the KL divergence [13]. Combining with (61) leads to (37).

A.7 Proof of lemma 2

Proof: Consider a Gaussian random vector $\mathbf{X} \in \mathbb{R}^d$ such that $P_{\mathbf{X}|Y}(\mathbf{x}|i) = \mathcal{G}(\mathbf{x}, \mu_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x})$, and define $\mathbf{X}_j = \pi_j^d(\mathbf{X})$. Since π_j^d is a linear transformation, $\mathbf{X}_j$ is also Gaussian distributed. Therefore, $P_{\mathbf{X}_j}(\mathbf{x})$ is uniquely determined by its mean and covariance. Using the well known relationships

$$\mu_{\mathbf{x}\,j} = \mathbf{\Pi}_j^d \mu_\mathbf{x}, \quad \text{and} \quad \mathbf{\Sigma}_{\mathbf{x}\,j} = \mathbf{\Pi}_j^d \mathbf{\Sigma}_\mathbf{x} (\mathbf{\Pi}_j^d)^T,$$

it follows that

$$P_{\mathbf{X}_j|Y}(\mathbf{x}|i) = \mathcal{G}(\mathbf{x}, \mathbf{\Pi}_j^d \mu_\mathbf{x}, \mathbf{\Pi}_j^d \mathbf{\Sigma}_\mathbf{x} (\mathbf{\Pi}_j^d)^T).$$

Applying this result to each of the C components of (41) we obtain (42).

References

[1] J. Bach. The Virage Image Search Engine: An open framework for image management. In *SPIE Storage and Retrieval for Image and Video Databases, San Jose, California*, 1996.

[2] A. Bell and T. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3328, December 1997.

[3] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color-and texture-based image segmentation using EM and its application to content-based image retrieval. In *International Conference on Computer Vision, Bombay, India*, pages 675–682, 1998.

[4] J. Bergen and E. Adelson. Early Vision and Texture Perception. *Nature*, 333(6171):363–364, 1988.

[5] J. Bergen and M. Landy. Computational Modeling of Visual Texture Segregation. In M. Landy and J. Movshon, editors, *Computational Models of Visual Processing*. MIT Press, 1991.

[6] J. De Bonet and P. Viola. Structure Driven Image Database Retrieval. In *Neural Information Processing Systems, Denver, Colorado*, volume 10, 1997.

[7] J. De Bonet, P. Viola, and J. Fisher. Flexible Histograms: A Multiresolution Target Discrimination Model. In E. G. Zelnio, editor, *Proceedings of SPIE*, volume 3370-12, 1998.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman & Hall, 1993.

[9] T. Caelli. A Brief Overview of Texture Processing in Machine Vision. In T. Papathomas, editor, *Early Vision and Beyond*, chapter 8. MIT Press, 1996.

[10] J. Cardoso. Blind Signal Separation: Statistical Principles. *Proceedings of the IEEE*, 90(8):2009–20026, October 1998.

[11] D. Comaniciu, P. Meer, K. Xu, and D. Tyler. Retrieval Performance Improvement through Low Rank Corrections. In *Workshop in Content-based Access to Image and Video Libraries, Fort Collins, Colorado*, pages 50–54, 1999.

[12] P. Comon. Independent Component Analysis, A New concept? *Signal Processing*, 36:287–314, 1994.

[13] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[14] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from Incomplete Data via the EM Algorithm. *J. of the Royal Statistical Society*, B-39, 1977.

[15] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[16] M. Do and M. Vetterli. Wavelet-based Testure Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance. *IEEE Trans. on Image Processing*, Vol. 11(2):146–158, February 2002.

[17] D. Dunn and W. Higgins. Optimal Gabor Filters for Texture Segmentation. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 7(4), July 1995.

[18] Y. Ephraim, A. Denbo, and L. Rabiner. A Minimum Discrimination Information Approach for Hidden Markov Modeling. *IEEE Trans. on Information Theory*, 35(5):1001–1013, September 1989.

[19] Y. Ephraim, H. Lev-Ari, and R. Gray. Asymptotic Minimum Discrimination Information Measure for Asymptotically Weakly Stationary Processes. *IEEE Trans. on Information Theory*, 34(5):1033–1040, September 1988.

[20] D. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, January 1989.

[21] I. Fogel and D. Sagi. Gabor Filters as Texture Discriminators. *Biol. Cybern.*, 61:103–113, 1989.

[22] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[23] B. Funt and G. Finlayson. Color Constant Color Indexing. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 17(5):522–529, May 1995.

[24] W. Gardner and B. Rao. Theoretical Analysis of the High-Rate Vector Quantization of LPC Parameters. *IEEE Trans. Speech and Audio Processing*, 3(5):367–376, September 1995.

[25] A. Gersho and R. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Press, 1992.

[26] T. Gevers and A. Smeulders. PickToSeek: Combining Color and Shape Invariant Features for Image Retrieval. *IEEE Trans. on Image Processing*, 9(1):102–119, January 2000.

[27] R. Gray A. Gray, G. Rebolledo, and J. Shore. Rate-Distortion Speech Coding with a Minimum Discrimination Information Distortion Measure. *IEEE Trans. on Information Theory*, vol. IT-27:708–721, November 1981.

[28] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Spatial Color Indexing and Applications. *Int. Journal of Computer Vision*, 35(3):245–268, December 1999.

[29] D. Hubel and T. Wiesel. Brain Mechanisms of Vision. *Scientific American*, September 1979.

[30] A. Hyvarinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13:411–430, 2000.

[31] G. Iyengar and A. Lippman. Clustering Images Using Relative Entropy for Efficient Retrieval. In *International workshop on Very Low Bitrate Video Coding, Urbana, Illinois*, 1998.

[32] N. Jayant and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, 1984.

[33] S. Kullback. *Information Theory and Statistics*. Dover Publications, New York, 1968.

[34] M. Kupperman. Probabilities of Hypothesis and Information-Statistics in Sampling from Exponential-Class Populations. *Annals of Mathematical Statistics*, 29:571–574, 1958.

[35] T. Leung and J. Malik. Representing and Recognizing the Visual Appearence of Materials using Three-dimensional Textons. *Int. Journal of Computer Vision*, Vol. 43(1):29–44, June 2001.

[36] H. Lev-Ari, S. Parker, and T. Kailath. Multidimensional Maximum-Entropy Covariance Extension. *IEEE Trans. on Information Theory*, 35(3):497–508, May 1988.

[37] J. Li and A. Barron. Mixture Density Estimation. In *Neural Information Processing Systems, Denver, Colorado*, 1999.

[38] J. Li, N. Chadda, and R. Gray. Asymptotic Performance of Vector Quantizers with a Perceptual Distortion Measure. *IEEE Trans. on Information Theory*, 45(4):1082–1091, May 1999.

[39] Q. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, 1999.

[40] F. Liu and R. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 18(3):722–733, July 1996.

[41] W. Ma and H. Zhang. Benchmarking of Image Features for Content-based Retrieval. In 32^{nd} *Asilomar Conference on Signals, Systems, and Computers, Asilomar, California*, 1998.

[42] J. Malik and P. Perona. Preattentive Texture Discrimination with Early Vision Mechanisms. *Journal of the Optical Society of America*, 7(5):923–932, May 1990.

[43] S. Mallat. A Theory for Multiresolution Signal Decomposition: the Wavelet Representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 11:674–693, July 1989.

[44] B. Manjunath and W. Ma. Texture Features for Browsing and Retrieval of Image Data. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 18(8):837–842, August 1996.

[45] J. Mao and A. Jain. Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models. *Pattern Recognition*, 25(2):173–188, 1992.

[46] G. McLean. Vector Quantization for Texture Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 23(3):637–649, May/June 1993.

[47] B. Moghaddam, H. Bierman, and D. Margaritis. Defining Image Content with Multiple Regions-of-Interest. In *Workshop in Content-based Access to Image and Video Libraries, Fort Collins, Colorado*, pages 89–93, 1999.

[48] H. Murase and S. Nayar. Visual Learning and Recognition of 3-D Objects from Appearence. *International Journal of Computer Vision*, 14:5–24, 1995.

[49] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, and G. Taubin. The QBIC project: Querying images by content using color, texture, and shape. In *SPIE Storage and Retrieval for Image and Video Databases, San Jose, California*, pages 173–181, 1993.

[50] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[51] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.

[52] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *ACM Intl. Multimedia Conference*, pages 65–73. ACM, Nov. 1996.

[53] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based Manipulation of Image Databases. *Int. Journal of Computer Vision*, Vol. 18(3):233–254, June 1996.

[54] R. Picard, T. Kabir, and F. Liu. Real-time Recognition with the entire Brodatz Texture Database. In *Proc. IEEE Conf. on Computer Vision, New York*, 1993.

[55] M. Porat and Y. Zeevi. Localized Texture Processing in Vision: Analysis and Synthesis in the Gaborian Space. *IEEE Trans. on Biomedical Engineering*, 36(1):115–129, January 1989.

[56] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical Evaluation of Dissimilarity Measures for Color and Texture. In *International Conference on Computer Vision, Korfu, Greece*, pages 1165–1173, 1999.

[57] T. Reed and J. Hans du Buf. A Review of Recent Texture Segmentation and Feature Extraction Techniques. *Computer Vision, Graphics, and Image Processing*, Vol. 57, May 1993.

[58] Y. Rui, T. Huang, and S.-F. Chang. Image Retrieval: Current Techniques, Promising Directions and Open Issues. *Journal of Visual Communication and Image Representation*, 10:39–62, March 1999.

[59] Yong Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998.

[60] D. Sagi. The Psychophysics of Texture Segmentation. In T. Papathomas, editor, *Early Vision and Beyond*, chapter 7. MIT Press, 1996.

[61] B. Schiele and J. Crowley. Recognition without Correspondence using Multidimensional Receptive Field Histograms. *International Journal of Computer Vision*, 36(1):31–50, January 2000.

[62] C. Schmid and R. Mohr. Local Greyvalue Invariants for Image Retrieval. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 19(5):530–535, May 1997.

[63] J. Simonoff. *Smoothing Methods in Statistics*. Springer-Verlag, 1996.

[64] J. Smith. *Integrated Spatial and Feature Image Systems: Retrieval, Compression and Analysis*. PhD thesis, Columbia University, 1997.

[65] M. Stricker and A. Dimai. Color Indexing with Weak Spatial Constraints. In *SPIE Storage and Retrieval for Image and Video Databases, San Jose, California*, volume 2670, pages 29–40, 1996.

[66] M. Stricker and M. Orengo. Similarity of Color Images. In *SPIE Storage and Retrieval for Image and Video Databases, San Jose, California*, 1995.

[67] A. Sutter, J. Beck, and N. Graham. Contrast and Spatial Variables in Texture Segregation: testing a simple spatial-frequency channels model. *Perceptual Psychophysics*, 46:312–332, 1989.

[68] M. Swain and D. Ballard. Color Indexing. *International Journal of Computer Vision*, Vol. 7(1):11–32, 1991.

[69] H. Tamura, S. Mori, and T. Yamawaki. Texture Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 6:460–473, 1978.

[70] L. Taycher, M. Cascia, and S. Sclaroff. Image Digestion and Relevance Feedback in the Image Rover WWW Search Engine. In *Visual, San Diego, California*, 1997.

[71] B. Thai and G. Healey. Spatial Filter Selection for Illumination-Invariant Color Texture Discrimination. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado*, 1999.

[72] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley, 1985.

[73] M. Tuceryan and A. Jain. Texture Analysis. In C. Chen, L. Pau, and P. Wang, editors, *The Handbook of Pattern Recognition and Computer Vision*, chapter 11. World Scientific Publishing Co., 1992.

[74] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3, 1991.

[75] K. Valkealahti and E. Oja. Reduced Multidimensional Co-Occurrence Histograms in Texture Classification. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 20(1):90–94, January 1998.

[76] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.

[77] N. Vasconcelos. *Bayesian Models for Visual Information Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[78] N. Vasconcelos and A. Lippman. Library-based Coding: a Representation for Efficient Video Compression and Retrieval. In *Proc. Data Compression Conference, Snowbird, Utah*, 1997.

[79] N. Vasconcelos and A. Lippman. Learning from user feedback in image retrieval systems. In *Neural Information Processing Systems, Denver, Colorado*, 1999.

[80] N. Vasconcelos and A. Lippman. Feature Representations fro Image Retrieval: Beyond the Color Histogram. In *Proc. Int. Conf. on Multimedia and Expo, New York*, 2000.

[81] N. Vasconcelos and A. Lippman. Learning Over Multiple Temporal Scales in Image Databases. In *Proc. European Conference on Computer Vision, Dublin, Ireland*, 2000.

CRL 2002/04
April 2002

Decision-theoretic Image Retrieval with Embedded Multi-resolution Mixtures

Nuno Vasconcelos