Minimum Probability of Error Image Retrieval

Nuno Vasconcelos, *Member, IEEE*

***Abstract*—We address the design of optimal architectures for image retrieval from large databases. Minimum probability of error (MPE) is adopted as the optimality criterion and retrieval formulated as a problem of statistical classification. The probability of retrieval error is lower- and upper-bounded by functions of the Bayes and density estimation errors, and the impact of the components of the retrieval architecture (namely, the feature transformation and density estimation) on these bounds is characterized. This characterization suggests interpreting the search for the MPE feature set as the search for the minimum of the convex hull of a collection of curves of probability of error versus feature space dimension. A new algorithm for MPE feature design, based on a dictionary of empirical feature sets and the wrapper model for feature selection, is proposed. It is shown that, unlike traditional feature selection techniques, this algorithm scales to problems containing large numbers of classes. Experimental evaluation reveals that the MPE architecture is at least as good as popular empirical solutions on the narrow domains where these perform best but significantly outperforms them outside these domains.**

***Index Terms*—Bayesian methods, color and texture, expectation–maximization, feature selection, image retrieval, image similarity, minimum probability of error, mixture models, multiresolution, optimal retrieval systems, wrapper methods.**

I. Introduction

GIVEN its dependence on text-based data-structures, existing database technology faces a new and difficult challenge with the ubiquitous emergence of multimedia databases. Because the automatic generation of natural language descriptors for multimedia signals is still beyond the reach of signal understanding technology, an entirely new database search paradigm has been advocated by various researchers over the last decade [1]–[5]. This new paradigm, which is commonly referred to as content-based retrieval, augments traditional text-based search with the ability to query by example: Users express their queries by providing examples of what they are looking for, and the target database items are retrieved by similarity to these user-provided examples.

While significant progress has been achieved, over the last decade, in various areas of the content-based retrieval problem, e.g., see [4]–[6] for extensive reviews of the literature, it is still not well understood how to design retrieval architectures that are optimal in an end-to-end sense, i.e., where all components are jointly optimized with respect to an overall optimality criterion or cost. Since, in the absence of a reason to favor certain types of errors, the natural goal of any retrieval system is to be correct as often as possible, it seems sensible to adopt, as optimality criterion, the minimization of the probability of retrieval error.[1] The retrieval problem is, in this way, formalized as one of supervised learning, or classification, and a vast body of existing knowledge in statistical learning becomes applicable to the design of optimal retrieval architectures. However, while statistical learning has been most successful in the context of relatively small classification problems (i.e., problems involving a small number of classes, typically two, and relatively small amounts of data per class), signal databases can easily contain thousands of classes and invariably generate large quantities of data per class. Due to this, many of the state-of-the-art solutions for problems such as classifier design or feature selection do not scale well enough to be applicable in the retrieval context.

In this work, we study the design of architectures for the evaluation of image similarity that are scalable and optimal with respect to the joint design of *minimum probability of error* (MPE) similarity functions, feature spaces, and density models. We consider lower and upper bounds on the probability of error of a retrieval system and study their dependence on these elements. The resulting theoretical characterization suggests interpretation of a feature transformation as a curve of probability of error versus feature space dimension, and the MPE solution as the minimum of the convex hull of all such curves. The analytical derivation of this minimum is, however, a difficult problem, because it depends on the particular classifier implementation. We propose a new algorithmic solution that combines the wrapper model for automated feature selection (FS) [7], [8] and prior knowledge about what are good features for various image domains.

We start from a dictionary of empirical feature transformations, i.e., transformations that have consistently met the challenge of extensive evaluation in specific image domains and use cross-validation to select the best feature subset for the target database. We then show that the wrapper approach to feature subset selection can be performed efficiently when all densities are modeled as Gauss mixtures and all feature transformations are linear and invertible. More precisely, we derive efficient recursive procedures for computing both model parameters and query similarity scores over sequences of embedded subspaces of the different feature transforms in the dictionary. It is shown that for databases with large numbers of image classes, the resulting MPE-retrieval architecture has complexity that is equivalent to that of a suboptimal retrieval system based on an arbitrarily chosen feature space of the same dimension.

Manuscript received July 2, 2003; revised March 2, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Meir Feder.

The author is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, 92093-0407 USA (e-mail: nuno@ece.ucsd.edu).

Digital Object Identifier 10.1109/TSP.2004.831125

[1]This formulation can also be easily extended to the case where there is a preference for certain types of errors, even though the issue is not addressed in this work.

Experimental evaluation on an image collection, containing diverse types of imagery and a large number of image classes, supports three main conclusions. First, the MPE feature space for a given database can be dramatically superior, in terms of probability of error, to an arbitrarily chosen space. This holds even when the latter is chosen from the same dictionary as the former or when the two are subspaces of the same feature space. Second, different feature transforms perform best on different databases, and a given feature transformation can have significant variations in probability of error when applied to different types of imagery. Third, the MPE-retrieval architecture performs at least as well as the empirical methods that are most popular in the retrieval literature on the databases containing imagery of the narrow domains for which they were proposed. On the other hand, for databases containing generic imagery, MPE retrieval achieves significant gains over these narrow-scope methods.

The paper is organized as follows. In Section II, we formulate the retrieval problem as one of supervised learning and review relevant results from learning theory. Section III addresses the issue of MPE image representation. The probability of error of a retrieval system is shown to be bounded by two functions of this representation, and the dependence of these bounds on the dimension of the space is characterized. The new MPE retrieval architecture is proposed in Section IV, where we discuss the merits of feature subset selection and derive the efficient wrapper algorithm for its implementation. Finally, experimental results are presented in Section V.

II. MPE Retrieval Systems

A retrieval system is a mapping

$$g : \mathcal{X} \rightarrow \mathcal{Y} = \{1, \ldots, M\}$$

from a feature space $\mathcal{X}$ to the index set $\mathcal{Y}$ of the M classes in the database. The retrieval system is optimal, under some suitable cost, if $\mathcal{X}$ and the similarity function $g(\cdot)$ are jointly optimized with respect to that cost. In this work, we adopt the minimization of the probability of retrieval error as the goal for this optimization.

Definition 1: An MPE retrieval system is the mapping

$$g^* : \mathcal{X} \rightarrow \mathcal{Y}$$

that, for all $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, minimizes

$$P_{\mathbf{X},Y}\left(g(\mathbf{x}) \neq y\right)$$

where $\mathbf{X}$ is the random variable from which the feature vector $\mathbf{x}$ is drawn, and Y is the random variable that assigns $\mathbf{x}$ to its database class.

It follows from this definition that an MPE retrieval system is an example of the classical Bayes classification problem [9]–[12]. While this makes various known classification results applicable to the retrieval problem, it should be pointed out that a retrieval system is not a standard classifier. Because, in the retrieval problem, the ultimate set of class-labels (relevant versus irrelevant to the query) is user and query dependent, these class labels are not known in advance of the classification operation. Hence, it is impossible to train a classifier that determines what is relevant to the user. An alternative solution, that we adopt in this work, is to assume that the class structure is an intrinsic property of the database. In particular, we assume that each image in the database is a class by itself. This is a solution of least commitment that enables the treatment of the problem in the traditional classification framework. It is also in line with various previous formulations of the problem, that by posing retrieval as some form of image matching (or matching of image descriptors such as histogram, feature vectors, etc.), [1], [2], [13]–[23] implicitly adopt the same strategy. The formulation, and all the algorithms presented in this work, are equally valid when the images are grouped according to some other predefined class structure (see, e.g., [24]). Finally, the ranking of images according to relevance/irrelevance to a particular user can still be implemented, for each query, with recourse to relevance feedback algorithms, e.g., as discussed in [25]. In this paper, we concentrate on the theoretical and algorithmic aspects of the basic classifier architecture.

A. MPE Classifiers

We start with a brief review of some known results on MPE classification; see, e.g., [12] for proofs.

Theorem 1: Consider an M-ary classifier $g(\cdot)$ of feature vectors $\mathbf{x}$ drawn from a random variable $\mathbf{X}$ in a feature space $\mathcal{X}$. The probability of error of $g(\cdot)$ is lower bounded by the *Bayes error*

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

where $E_{\mathbf{x}}$ means expectation with respect to $P_{\mathbf{X}}(\mathbf{x})$, and $P_{Y|\mathbf{X}}(i|\mathbf{x})$ is the posterior probability of class i given $\mathbf{x}$.

It follows that the Bayes error (BE) is the fundamental limit to the performance of MPE retrieval systems. It is also well known that this bound is tight, in the sense that there is always a classifier that achieves it.

Theorem 2: Given a feature space $\mathcal{X}$ and query feature vector $\mathbf{x}$, the similarity function that minimizes the probability of retrieval error is the *Bayes* classifier

$$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}). \tag{2}$$

The probability of error of the Bayes classifier is the BE.

The two theorems establish the Bayes classifier as the optimal similarity function for MPE retrieval systems. This, however, assumes the knowledge of the class-posterior probabilities $P_{Y|\mathbf{X}}(i|\mathbf{x})$, which are usually not available and must be estimated from a finite training sample. One popular solution is to rely on Bayes' rule

$$g^*(\mathbf{x}) = \arg\max_i P_{\mathbf{X}|Y}(\mathbf{x}|i)P_Y(i) \tag{3}$$

where $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ is the probability density for the feature vectors drawn from the ith class, and $P_Y(i)$ is the prior probability for that class, and approximate the optimal decision rule by

$$g(\mathbf{x}) = \arg\max_i \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\hat{p}_Y(i) \tag{4}$$

where $\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)$ and $\hat{p}_Y(i)$ are estimates of the quantities in (3). If there is no *a priori* reason to favor any of the image classes in the database, it is acceptable to assume that the class priors $P_Y(i)$ are known and uniform, i.e., $\hat{p}_Y(i) = P_Y(i) = 1/M$. This leads to a decision function that depends only on estimates for the class-conditional densities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$

$$g(\mathbf{x}) = \arg\max_i \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i). \tag{5}$$

Because these class-conditional densities can be estimated independently for each class, the overall training complexity scales linearly in the number of classes, making this classifier architecture particularly appealing for problems such as image retrieval or speech recognition [26], where that number is large. On the other hand, it should be emphasized that (5) is optimal, in the MPE sense, only insofar as the probability estimates are error-free. This is a requirement that is never met in practice, where density estimates are based on a finite data sample and, therefore, have nonzero variance. In fact, when the feature space $\mathcal{X}$ is high dimensional, the density estimation error can be substantial.

The overall probability of retrieval error therefore has two components: 1) the BE that, as shown by (1), only depends on $\mathcal{X}$ and 2) the density estimation error that depends both on $\mathcal{X}$ and the procedure used to obtain the density estimates. In Section III, we study the problem of achieving the optimal balance, in the MPE sense, between the two sources of error. For now, we assume the existence of a space of observations $\mathcal{Z}$, e.g., the space of $n \times n$ image blocks, and investigate the benefits of introducing a feature transformation $T : \mathcal{Z} \to \mathcal{X}$. The following theorem provides some insight on how the selection of $\mathcal{X}$ affects the BE.

Theorem 3: Given a retrieval system with observation space $\mathcal{Z}$ and a feature transformation

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

then

$$L^*_{\mathcal{X}} \geq L^*_{\mathcal{Z}} \tag{6}$$

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

Proof: See [12] for the case of two-class problems and [27] for an extension to multiple classes. ■

III. MPE Signal Representation

The design of MPE retrieval systems requires a good understanding of how the selection of both the feature space and the probability models used for density estimation affect the probability of error. This question can be decomposed into two simpler problems: 1) how the representation components affect the Bayes and estimation errors and 2) the impact of these errors on the probability of error. We start by addressing the second question.

A. Impact of Bayes and Estimation Errors on the Probability of Error

We have already seen that the BE is a lower bound on the probability of error. Furthermore, when the density estimation error is null, (5) is equivalent to the Bayes classifier, and therefore, the probability of error is equal to the BE. Intuitively, the impact of poor density estimates should be to increase the difference between these two quantities. This intuition is quantified by the following theorem.

Theorem 4: Consider a retrieval problem with equiprobable classes $P_Y(i) = 1/M, \forall i$, a feature space $\mathcal{X}$, unknown class conditional densities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, and the decision function of (5). For such a retrieval problem, the difference between the probability of error and the BE is upper bounded by

$$P_{\mathbf{X},Y}\left(g(\mathbf{X}) \neq Y\right) - L^*_{\mathcal{X}} \leq \Delta_{g,\mathcal{X}} \tag{7}$$

where

$$\Delta_{g,\mathcal{X}} = \frac{\sqrt{2\ln 2}}{M} \sum_i \sqrt{\mathrm{KL}\left[P_{\mathbf{X}|Y}(\mathbf{x}|i) \| \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\right]} \tag{8}$$

is the density estimation error, and

$$\mathrm{KL}\left[P_{\mathbf{X}|Y}(\mathbf{x}|i) \| \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\right] = \int P_{\mathbf{X}|Y}(\mathbf{x}|i) \log \frac{P_{\mathbf{X}|Y}(\mathbf{x}|i)}{\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)} d\mathbf{x} \tag{9}$$

is the Kullback–Leibler (KL) divergence [28] between the true $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ and estimated $\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)$ densities for class i.

Proof: See Appendix A. ■

We note that this is a bound on the distance between the actual probability of error and the Bayes error and is substantially different from various bounds available in the information theoretic literature (see, e.g., [29] and [30]) that relate BE to the KL divergence between class densities. In these bounds, the KL divergence appears as a measure of discrimination, and the bounds formalize the intuition that the BE decreases when the separation between class-conditional densities increases, i.e., with the increase of the KL divergence between classes. In the theorem above, the KL divergence does not take the role of a measure of discrimination, but instead, it appears as a measure of density estimation error. In particular, instead of the KL divergence between class-conditional densities, (7) is a function of the KL divergence between the true class-conditional densities and their estimates. It quantifies the statement that the probability of error is lower bounded by the BE and upper bounded by the sum of the BE and the estimation error.

B. Impact of the Representation Components on the Bayes and Estimation Errors

The components of signal representation affect the Bayes and estimation errors in very distinct ways. Since the BE only depends on the true densities and not their estimates, the only impact of the density model is on the estimation error. The relationships between these two quantities have been extensively studied in the statistics literature and are fairly well understood [11], [31]–[33]. We will not review them here but will simply select, in Section IV, a model that is well suited to the retrieval problem. For now, we concentrate on the dependence of the Bayes and estimation errors on the feature transformation.

1) Embedded Feature Spaces: We start by considering sequences of nested vector spaces of increasing dimension, which are also known as sequences of embedded vector spaces [34].

Definition 2: 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 one-to-one mappings

$$\epsilon_i : \mathcal{X}_i \to \mathcal{X}'_{i+1}, \quad i = 1, \ldots, d-1 \tag{10}$$

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

The concept of a sequence of embedded feature spaces plays a central role in this work for two main reasons. The first is theoretical and will be addressed in the remainder of this section. It motivates the search for the MPE feature transformation as the search for the minimum of the convex hull of a series of curves of probability of error, which is the basis for the feature selection algorithms introduced in Section IV. The second is algorithmic and will be discussed in Section IV. It resides in the fact that, for certain parametric probabilistic models, parameter estimates in all the embedded subspaces of a given vector space can be obtained *in closed form* once the parameter estimates are available for that space. This enables very large computational savings over the alternative of estimating parameters in all subspaces and makes the search for the minimum of the convex hull of probability of error feasible from a computational point of view.

2) Tradeoff Between Bayes and Estimation Errors: We start with the theoretical motivation.

Theorem 5: Let

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

be a linear feature transformation, and

$$\pi_m^n : \mathbb{R}^n \to \mathbb{R}^m \tag{11}$$

where $\pi_m^n(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n) = (x_1, \ldots, x_m)$, which is the projection of the Euclidean space along the coordinate axes. Then

$$\mathcal{X}_i = \pi_i^d(\mathcal{X}), \quad i = 1, \ldots, d-1 \tag{12}$$

is a sequence of embedded feature spaces such that

$$L^*_{\mathcal{X}_{i+1}} \leq L^*_{\mathcal{X}_i}. \tag{13}$$

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

$$\mathbf{X}_i = \pi_i^d(\mathbf{X}), \quad i = 1, \ldots, d \tag{14}$$

and $\{g_i(\mathbf{x})\}_{i=1}^d$ is a sequence of decision functions

$$g_i(\mathbf{x}) = \arg\max_k \hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}|k) \tag{15}$$

then

$$\Delta_{g_{i+1}, \mathcal{X}_{i+1}} \geq \Delta_{g_i, \mathcal{X}_i}. \tag{16}$$

Proof: See Appendix B. ■

The theorem shows that any *linear feature transformation originates a sequence of embedded vector spaces with monotonically decreasing Bayes error and monotonically increasing estimation error*. It follows that it is impossible to find a feature transformation that can minimize the Bayes and estimation errors simultaneously. On one hand, given a feature space $\mathcal{X}$, it is possible to find a subspace where density estimates are more accurate. On the other hand, the projection onto this subspace will increase the BE. The practical result is that for any feature transform used in a retrieval system, there is a need to reach a compromise between the two sources of error. This is illustrated by the left plot of Fig. 1, 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 BE. For higher dimensional subspaces, the decrease in BE 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 BE ceases to be dominant. Different feature transforms will originate different curves, and to achieve optimality, in the MPE sense, a retrieval system must operate on the minimum of the convex hull of all these curves. This is illustrated by the right plot of Fig. 1 and motivates the MPE retrieval architecture introduced in Section IV.

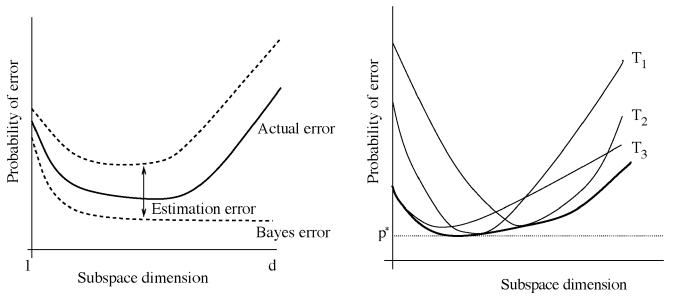

Fig. 1. (Left) Upper bound, lower bound, and probability of error as a function of subspace dimension. (Right) Curves of probability of error associated with three feature transformations. The convex hull is shown in dark, and its minimum value is p^*.

3) Relationship to Structural Risk Minimization: Before presenting the details of this architecture, we note that the combination of the construct of a sequence of embedded spaces (Theorem 5) and the search for the MPE embedded subspace of a given feature space are an implementation of the principle of structural risk minimization (SRM) for the design of learning machines [35]. The SRM principle relies on a nested subset of decision functions $S_1 \subset S_2 \subset S_3 \ldots$ and chooses the decision function that minimizes the classification error on a training set (usually referred to as *empirical risk*) from the subset that provides the best guarantees in terms of the actual probability of error. These guarantees are given in the form of bounds of the type

$$P_{\mathbf{X},Y}\left(g_i(\mathbf{X}) \neq Y\right) \leq \hat{P}_{\mathbf{X},Y}\left(g_i(\mathbf{X}) \neq Y\right) + \Phi(S_i) \tag{17}$$

where $g_i \in S_i$, $P(.)$ denotes the true probability, $\hat{P}(.)$ the empirical estimate obtained from a training set of size n, and $\Phi(S_i)$ is confidence interval, which is usually closely related to an upper bound of the supremum of the error within the class S_i. While various bounds are possible (see, e.g., the discussion in [36]), the most popular are based on the Vapnik–Chervonenkis

(VC) dimension of the subset S_i [35]. In this case, the SRM principle trades the quality of the approximation of the training data with the complexity of the approximating function, as measured by the VC dimension; for larger i, while the minimum $\hat{P}(g_i(\mathbf{X}) \neq Y)$ decreases, $\Phi(S_i)$ increases. The comparison of (17) with (7) reveals a similarity of roles for the terms 1) $\Phi(S_i)$ and $L^*_{\mathcal{X}_i}$ and 2) $\hat{P}_{\mathbf{X},Y}(g_i(\mathbf{X}) \neq Y)$ and $\Delta_{g_i,\mathcal{X}_i}$, which make the search for the MPE subspace of $\mathcal{X}$ a form of SRM. The attractive property of Theorem 5 is that, unlike for example the bounds based on the VC dimension, it depends directly on the dimension of the space. It therefore makes a strong argument for relying on the dimensionality of the space as the parameter that controls the actual probability of error. In the following section, we introduce a retrieval architecture based on this principle.

IV. Retrieval Architecture

The implementation of the MPE retrieval architecture requires the design of a density estimation and a feature selection module.

A. Density Estimation

Density estimates are typically obtained by fitting a parametric model to a training sample, usually by finding the set of parameters of maximum likelihood with respect to the training sample. We adopt this framework and rely on the popular family of Gauss mixture densities [11], [37] to obtain all parameter estimates.

Definition 3: A Gauss mixture is a density of the form

$$P_{\mathbf{X}}(\mathbf{x}) = \sum_{i=1}^{C} \lambda_i \mathcal{G}(\mathbf{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{18}$$

where

$$\mathcal{G}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}}} \tag{19}$$

is a Gaussian component of mean μ and nonsingular covariance $\boldsymbol{\Sigma}$

$$\|\mathbf{x} - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{20}$$

and λ_i is the probability of the ith component $(\sum_i \lambda_i = 1)$.

The Gauss mixture is appealing as a probabilistic model for retrieval since it provides a description of the true density that is compact enough to lead to a similarity function with tractable complexity, is tractable in high-dimensional spaces, and can approximate multimodal densities.

B. Feature Extraction and Selection

We start with a brief review of the predominant strategies for finding optimal features. Feature extraction (FE) techniques pose the problem as that of finding the optimal feature transformation T^* directly from training data. This is usually done through an optimization procedure, e.g., gradient descent on the space of $m \times n$ matrices when T is a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$. FS techniques assume that the optimal features are available but included in a larger set that is highly redundant. Given a collection of n features, the problem is to select the best subset of cardinality $m < n$.

Both strategies suffer from significant limitations. Since the space of transform matrices is high-dimensional, the optimization problem of FE is usually quite difficult to solve, in terms of the computation required, the ability to escape local minima, and the availability of sufficient training data to guarantee generalization. All these limitations are magnified by classification problems containing large numbers of classes, such as retrieval. Globally optimal FS is equally intractable since it requires the evaluation of the cost for all possible m-subsets of the n features [40]. The most popular alternative is to rely on greedy, or sequential, search procedures [38] that attain locally optimal solutions. These solutions are frequently better than the local minima in which FE tends to get trapped. Nevertheless, FS solutions can also be quite suboptimal. For example, highly discriminant linear combinations of feature are, in general, quite difficult to detect. Finally, the complexity of FS is usually intractable in the context of large-scale classification problems, even when sequential search methods are used [39].

One solution that is popular in the retrieval literature is to disregard optimality and adopt what we denote by *empirical* feature transformations. These are transformations that 1) at some intuitive level have good properties for the classification task at hand and 2) have met the challenge of extensive empirical evaluation in certain image domains. For example, the coefficients of an autoregressive model have been shown to perform well on texture databases [41], color histograms have been quite successful in various object recognition tasks [13], principal component analysis representations are widely used for face recognition [42], edge-based features have been proposed for shape databases [17], and Haar wavelets have worked well for pedestrian detection [43]. This strategy eliminates the complexity of determining the optimal set of features, and the resulting classifiers usually work quite well in the database domains for which they were designed. The main limitation is, however, an obvious lack of generalization (e.g., auto-regressive models tend not to work well on face databases, color histograms fail on texture databases, and so forth).

In this work, we propose an alternative strategy that combines aspects from the empirical and FS approaches and is motivated by Fig. 1. Since FS performs a search over the set of all combinations of features, it solves the discrete optimization problem of selecting the curve, among those associated with all the combinations, that touches the convex hull of the probability of error in the point closest to p^*. However, because many feature combinations originate curves of probability of error that never come close to the convex hull, FS is highly inefficient: The optimizer has to sort through many poor feature groupings in order to be able to find the few ones that are real candidates for the best transformation. It would be desirable to restrict the search to those feature combinations that do come close to the convex hull.

The empirical strategy is a limiting case of this approach: Because empirical transformations achieve near-optimal performance on some database classes, their curves of probability of error must indeed be close to the convex hull in some set of points. However, this set is usually small, and the resulting re-

trieval system does not generalize well. Nevertheless, the fact that it is usually possible to find an empirical transformation that performs well for any type of database suggests that by taking the union of all such transformations, it should be possible to construct a feature set capable of achieving a probability of error that is close to the convex hull for most databases of practical interest. In fact, this should be possible without an exhaustive search within the set of empirical features since these already come grouped into feature transformations that perform well in different regions of the convex hull.

This observation suggests a strategy consisting of 1) the adoption of a dictionary of empirical feature transforms and 2) a search for the best transform for each target database. This strategy is significantly simpler than the search for the best arbitrary combination of features performed by traditional FS. If, for example, there are F transformations and each produces k features, the number of possible solutions decreases from

$$O\left(\frac{Fk!}{k!(F-1)k!}\right) > O\left(\frac{[(F-1)k]^k}{k!}\right) > O\left((F-1)^k\right)$$

to $O(F)$! We next investigate how to implement this strategy.

C. MPE Basis Selection

We start by recalling that for any invertible linear feature transformation $T : \mathcal{Z} \to \mathcal{X}$, it is possible to define an inverse, reconstruction, mapping

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

The columns of the associated matrix $\mathbf{A}$ are called basis functions, and $\mathbf{A}$ is the basis matrix for T. The rows of $\mathbf{T}$ are called the filter functions, or filters, of the transformation. They are the same as the basis functions when the transformation is orthonormal. Since there is a one-to-one mapping between invertible linear feature transformations and their bases, we will use the two terms indiscriminately.

Definition 4: A *bases dictionary* is a set $\mathcal{T} = \{T^{(1)}, \ldots T^{(F)}\}$ of invertible linear transformations. $T \in \mathcal{T}$ is the MPE basis in the dictionary for a given database if it achieves smaller probability of error in that database than all other bases in $\mathcal{T}$.

Clearly, the search for the MPE basis requires a strategy for evaluating the probability of error of a given basis. Since the latter depends, in nontrivial ways, on both the basis and the classifier, this requires cross-validation. Cross-validation is a well-established procedure in the statistical literature (see, e.g., [44]) that relies on the training set itself to evaluate classification performance. While many variations are possible, it always consists of removing a subset of examples (the cross-validation set) from the training set, designing a classifier with the remaining ones, and estimating the classification rate by that obtained on the cross-validation set. The process is usually repeated with different subsets of training examples as cross-validation sets.

The idea of using classification performance (measured with cross-validation) as the criterion for FS, by including the classifier in the FS loop, is a popular one in the machine learning literature, where the procedure is commonly referred to as the *wrapper model* for FS [7], [8]. However, the straightforward application of this model would add a significant computational burden to the design of a retrieval system since it requires redesigning the classifier and evaluating all cross-validation queries on all feature subspaces associated with all transformations in $\mathcal{T}$. In the remainder of this section, we show that an efficient implementation of the wrapper model exists when the retrieval architecture follows the probabilistic retrieval model, all densities are Gauss mixtures, and all feature transformations are linear and invertible. The starting point is the following property of the Gauss mixture.[2]

Property 1: If $\mathbf{T} : \mathcal{X} \to \mathcal{X}'$ is a linear feature transformation $\mathbf{X} \in \mathcal{X}$ and $\mathbf{X}' \in \mathcal{X}'$ are two random variables such that $\mathbf{X}$ is distributed according to (18) and $\mathbf{X}' = \mathbf{T}\mathbf{X}$, then

$$P_{\mathbf{X}'}(\mathbf{x}) = \sum_i \lambda_i \mathcal{G}(\mathbf{x}, \mathbf{T}\boldsymbol{\mu}_i, \mathbf{T}\boldsymbol{\Sigma}_i\mathbf{T}^T). \tag{21}$$

D. MPE Subspace Selection and Embedded Mixture Models

Consider a linear transformation $\mathbf{T}$ and associated feature space $\mathcal{X} \subset \mathbb{R}^d$. From Theorem 5, the sequence $\mathcal{X}_j = \pi_i^d(\mathcal{X})$ is a sequence of embedded subspaces of $\mathcal{X}$. As discussed in Section III, the characterization of the probability of error achievable with $\mathbf{T}$ requires the determination of its MPE subspace. Denoting by $\boldsymbol{\Pi}_j$ the projection matrix associated with π_j^d, i.e., $\boldsymbol{\Pi}_j = [\mathbf{I}_j, \mathbf{0}_{d-j}]$, where $\mathbf{I}_j$ is the identity matrix of order j and $\mathbf{0}_{d-j}$ the $j \times d-j$ zero matrix, it follows from the property above that if $\mathbf{X} \in \mathcal{X}$ is distributed according to (18), the random variables $\mathbf{X}_j = \pi_j^d(\mathbf{X})$ are distributed according to

$$P_{\mathbf{X}_j}(\mathbf{x}) = \sum_i \lambda_i \mathcal{G}\left(\mathbf{x}, \boldsymbol{\Pi}_j\boldsymbol{\mu}_i, \boldsymbol{\Pi}_j\boldsymbol{\Sigma}_i\boldsymbol{\Pi}_j^T\right). \tag{22}$$

The collection of densities in (22) is the family of *embedded mixture models* associated with $\mathbf{X}$. It has two properties of significant practical interest. The first is that once an estimate is available for $\{\lambda_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$, the parameters of $P_{\mathbf{X}_j}(\mathbf{x})$ can be obtained for any j by simply extracting the first j components of the mean vectors $\boldsymbol{\mu}_i$ and the upper-left $j \times j$ submatrix of the covariances $\boldsymbol{\Sigma}_i$. This implies that it is not necessary to repeat the density estimation for each of the subspace dimensions under consideration. Hence, the complexity of estimating all $P_{\mathbf{X}_j}(\mathbf{x})$ is the same as that of estimating $P_{\mathbf{X}}(\mathbf{x})$. The second is a similar result for the complexity of evaluating the queries in the cross-validation set. It is based on the fact that the complexity of (18) is dominated by the computation of $\|\mathbf{x} - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$ and $|\boldsymbol{\Sigma}|$.

Lemma 1: Consider the contribution to $P_{\mathbf{X}_j}(\boldsymbol{\Pi}_j\mathbf{x})$, $j = 1, \ldots, d$ of a mixture component with mean $\boldsymbol{\Pi}_j\boldsymbol{\mu}$ and covariance $\mathbf{S}_j = \boldsymbol{\Pi}_j\boldsymbol{\Sigma}\boldsymbol{\Pi}_j^T$. The terms $\mathcal{M}_j = \|\boldsymbol{\Pi}_j\mathbf{x} - \boldsymbol{\Pi}_j\boldsymbol{\mu}\|_{\mathbf{S}_j}$ and $\mathcal{D}_j = |\mathbf{S}_j|$ are given by the following recursion.

Initial conditions: $\mathcal{M}_1 = (x_1 - \mu_1)^2/\sigma_{1,1}$, $\mathcal{D}_1 = \mathbf{S}_1 = \sigma_{1,1}$.

Recursion:

$$\boldsymbol{\psi}_j^T = \left(\mathbf{u}_{j-1}^T\mathbf{S}_{j-1}^{-1}, -1\right) \tag{23}$$

$$p_j = -\left(\mathbf{u}_{j-1}^T, \sigma_{j,j}\right)\boldsymbol{\psi}_j \tag{24}$$

[2]This property follows trivially from the equivalent, and well-known, property for Gaussian random variables.

$$\mathbf{S}_j^{-1} = \Gamma\left(\mathbf{S}_{j-1}^{-1}\right) + \frac{1}{p_j}\boldsymbol{\psi}_j\boldsymbol{\psi}_j^T \tag{25}$$

$$\mathcal{M}_j = \mathcal{M}_{j-1} + \frac{\left(\boldsymbol{\psi}_j^T \boldsymbol{\Pi}_j \mathbf{d}\right)^2}{p_j} \tag{26}$$

$$\mathcal{D}_j = p_j \mathcal{D}_{j-1} \tag{27}$$

where $\Gamma(\cdot)$ is a mapping that adds to matrix $\cdot$ a row and a column (which become the last row and column, respectively) of zeros, $\mathbf{d} = \mathbf{x} - \boldsymbol{\mu}$, $\sigma_{i,j}$ is the (i,j)th element of $\boldsymbol{\Sigma}$, and $\mathbf{u}_{j-1} = (\sigma_{1,j}, \ldots, \sigma_{j-1,j})^T$ the vector containing the $j-1$ first elements of the jth column of $\boldsymbol{\Sigma}$. The complexity of evaluating all $\mathcal{M}_j$ and $\mathcal{D}_j$ is $O(d^3)$.

Proof: See Appendix C. ∎

It follows from this lemma that the complexity of evaluating all $P_{\mathbf{X}_j}(\boldsymbol{\Pi}_j\mathbf{x})$ is $O(d^3)$, and since this is also the cost of computing $\boldsymbol{\Sigma}^{-1}$, this complexity is the same as that required to compute $P_{\mathbf{X}}(\mathbf{x})$. Hence, the search for the MPE subspace of a query $\mathbf{x}$ does not impose any increase in complexity over the simple evaluation of the similarity score $P_{\mathbf{X}}(\mathbf{x})$ for that query. It should be noted that the lemma assumes the knowledge of the MPE ordering of the embedded subspaces, without which, a combinatorial search for the optimal subspace ordering would be required. While, in practice, the optimal subspace ordering is not known, an empirical ordering that works well in the target domain is usually available for any given empirical feature transformation. We will return to this topic in Section IV-F.

E. Selection of the Optimal Feature Transform

The following theorem extends the results of the previous section to the case where the goal is to find the MPE-subspace among all transformations in a basis dictionary.

Theorem 6: Consider a basis dictionary $\mathcal{T}$ and the random variable $\mathbf{X}^{(l)}$ associated with $\mathcal{X}^{(l)}$, which is the range space of the l^{th} feature transformation $T^{(l)}$. Let $\mathbf{X}^{(l)}$ be distributed according to a Gauss mixture of parameters $\{\lambda_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$. Then, for any $m \in \{1, \ldots, F\}$, the random variables $\mathbf{X}_j^{(m)} = \pi_j^d(\mathbf{X}^{(m)})$ are distributed according to a sequence of embedded Gauss mixtures of the form

$$P_{\mathbf{X}_j^{(m)}}(\mathbf{x}) = \sum_i \lambda_i \times \mathcal{G}\left(\mathbf{x}, \boldsymbol{\Pi}_j\mathbf{T}^{(l,m)}\mu_i, \boldsymbol{\Pi}_j\mathbf{T}^{(l,m)}\boldsymbol{\Sigma}_i\left(\boldsymbol{\Pi}_j\mathbf{T}^{(l,m)}\right)^T\right) \tag{28}$$

where $\mathbf{T}^{(l,m)} = \mathbf{T}^{(m)}(\mathbf{T}^{(l)})^{-1}$. Furthermore, the contribution to $P_{\mathbf{X}_j^{(m)}}(\boldsymbol{\Pi}_j\mathbf{T}^{(l,m)}\mathbf{x})$, $j = 1, \ldots, d$ of a mixture component of mean $\boldsymbol{\Pi}_j\mathbf{T}^{(l,m)}\boldsymbol{\mu}$ and covariance $\mathbf{S}_j^{(l,m)} = \boldsymbol{\Pi}_j\mathbf{T}^{(l,m)}\boldsymbol{\Sigma}(\boldsymbol{\Pi}_j\mathbf{T}^{(l,m)})^T$ can be computed recursively. Letting $\mathcal{M}_j^{(l,m)} = ||\boldsymbol{\Pi}_j\mathbf{T}^{(l,m)}\mathbf{x} - \boldsymbol{\Pi}_j\mathbf{T}^{(l,m)}\boldsymbol{\mu}||_{\mathbf{S}_j^{(l,m)}}$ and $\mathcal{D}_j^{(l,m)} = |\mathbf{S}_j^{(l,m)}|$, the recursion is as follows.

Initial conditions:

$$\mathcal{M}_1^{(l,m)} = \frac{\left(\boldsymbol{\Pi}_1\mathbf{T}^{(l,m)}(\mathbf{x}-\boldsymbol{\mu})\right)^2}{\mathbf{S}_1^{(l,m)}}$$

$$\mathcal{D}_1^{(l,m)} = \mathbf{S}_1^{(l,m)} = \sigma_{1,1}^{(l,m)}.$$

Recursion:

$$\left(\boldsymbol{\psi}_j^{(l,m)}\right)^T = \left[\left(\mathbf{u}_{j-1}^{(l,m)}\right)^T \left(\mathbf{S}_{j-1}^{(l,m)}\right)^{-1}, -1\right]$$

$$p_j^{(l,m)} = -\left[\left(\mathbf{u}_{j-1}^{(l,m)}\right)^T, \sigma_{j,j}^{(l,m)}\right]\boldsymbol{\psi}_j^{(l,m)}$$

$$\left(\mathbf{S}_j^{(l,m)}\right)^{-1} = \Gamma\left[\left(\mathbf{S}_{j-1}^{(l,m)}\right)^{-1}\right] + \frac{1}{p_j^{(l,m)}}\boldsymbol{\psi}_j^{(l,m)}\left(\boldsymbol{\psi}_j^{(l,m)}\right)^T$$

$$\mathcal{M}_j^{(l,m)} = \mathcal{M}_{j-1}^{(l,m)} + \frac{\left[\left(\boldsymbol{\psi}_j^{(l,m)}\right)^T \Pi_j\mathbf{T}^{(l,m)}\mathbf{d}\right]^2}{p_j^{(l,m)}}$$

$$\mathcal{D}_j^{(l,m)} = p_j^{(l,m)}\mathcal{D}_{j-1}^{(l,m)}$$

where $\Gamma(\cdot)$ is a mapping that adds to matrix $\cdot$ a row and a column (which become the last row and column, respectively) of zeros, $\mathbf{d} = \mathbf{x} - \boldsymbol{\mu}$, $\sigma_{i,j}^{(l,m)}$ is the $(i,j)^{th}$ element of $\mathbf{T}^{(l,m)}\boldsymbol{\Sigma}(\mathbf{T}^{(l,m)})^T$, and $\mathbf{u}_{j-1}^{(l,m)}$ the vector containing the $j-1$ first elements of the jth column of this matrix. Given l, the complexity of evaluating all $\mathcal{M}_j^{(l,m)}$ and $\mathcal{D}_j^{(l,m)}$ is $O(Fd^3)$.

Proof: See Appendix D. ∎

The procedure required to 1) map the parameters of a mixture component $\{\lambda, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ from a reference space $\mathcal{X}^{(l)}$ to the embedded subspaces of all other $\mathcal{X}^{(m)}$ associated with $\mathcal{T}$ and 2) recursively computing all terms required for evaluating the contribution of that component to the similarity score of a query $\mathbf{x} \in \mathcal{X}^{(l)}$ in all these subspaces is illustrated in Fig. 2. The basic building block is a downward projection of the parameters of the mixture component, followed by the recursion that propagates $\mathcal{M}_j^{(l,m)}$ and $\mathcal{D}_j^{(l,m)}$ from the lowest dimensional subspace to the full space. This building block is replicated for each or the feature transformations in $\mathcal{T}$.

Theorem 6 characterizes the computational cost of cross-validation in terms of both density estimation and query evaluation. With respect to the former, the theorem shows that given the parameters $\{\lambda_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ associated with one of the transforms in $\mathcal{T}$, it is straightforward to obtain the parameters associated with all other transforms in the dictionary. Hence, the complexity of estimating densities on all feature spaces is equal to that required to estimate them in only one space. This implies that designing all classifiers necessary for MPE feature subset selection by cross-validation does not require more computation than the suboptimal approach of designing a single classifier on an arbitrarily chosen feature space. With respect to the complexity of query evaluation, the theorem shows that the complexity of searching for the MPE subspace of the entire dictionary is equivalent to that of performing FS queries on a suboptimal system with an arbitrarily chosen feature space, where S is the total number of images that are used in the cross-validation set.

Overall, when both density estimation and the evaluation of the cross-validation queries are accounted for, the training time of the MPE architecture now proposed is equal to the sum of the training time required by a suboptimal retrieval system on an arbitrary feature space (t_1) plus the time required by such a system to evaluate FS queries (t_2). When the image database on which the retrieval system will operate has a large number of classes, the former (t_1) is dominant, and the training time is, therefore,

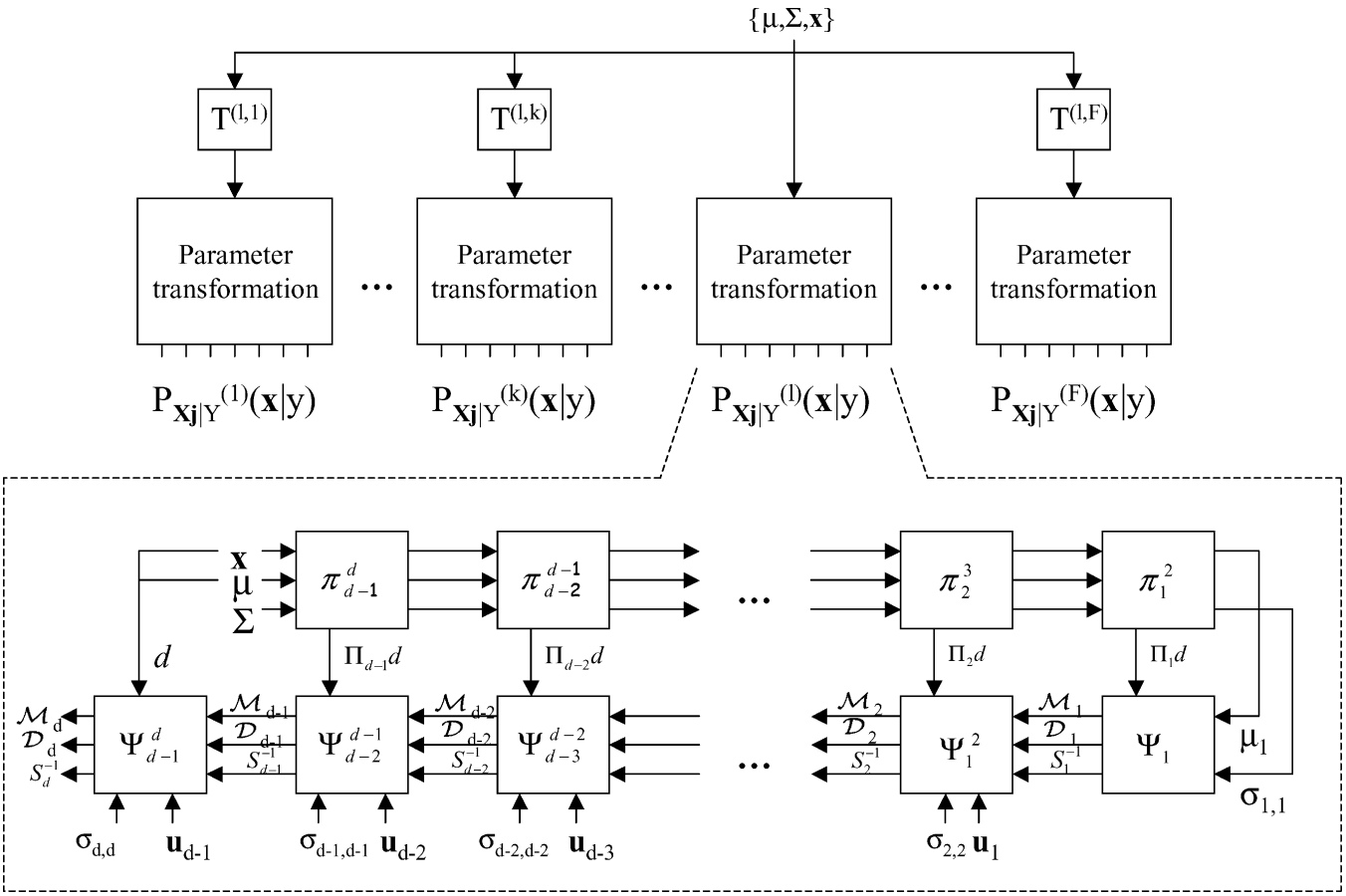

Fig. 2. Recursive algorithm for 1) propagating the parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ of a mixture component in a reference space $\mathcal{X}^{(l)}$ to all other features subspaces associated with $\mathcal{T}$ and 2) computing the components $\mathcal{D}_j$ and $\mathcal{M}_j$ of the similarity function on all feature subspaces. Ψ_{j-1}^{j} is the mapping that implements (23)–(27).

not substantially affected by cross-validation. Note that once the MPE-feature space is found, the complexity of query evaluation is exactly the same as that of a suboptimal retrieval system (since the other bases in the dictionary are no longer considered). Hence, there is no increase in retrieval time or in the time required by operations other than training, e.g., building indexes.

F. Embedded Multiresolution Mixture Models

One interesting question is whether the restriction to the set of linear transforms will significantly reduce the performance of the resulting MPE-retrieval systems. After all, empirical transforms such as the set of coefficients of a multiresolution simultaneous auto-regressive (MRSAR) model [45], which is quite popular in the texture retrieval literature [18], [22], [23], are not linear. While absolute conclusions can only be reached by experimental evaluation, we believe that the linear restriction is not necessarily a major limitation.

There are a few reasons for this. The first is that many of the empirical transformations that have been proposed in the literature are indeed linear. Examples include all wavelet decompositions, principal component analysis, Fourier transforms, Gabor functions, and various others. The second is the biological plausibility of linear transformations. Ever since the work of Hubel and Wiesel [46], it has been established that 1) human 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 approximated by a multiresolution representation localized in space and frequency, and several "biologically plausible" models of early vision are based on this principle [47]–[52]. More recently, it has been shown that filters remarkably similar to the receptive fields of cells found in V1 [53], [54] can be learned from training images by imposing requirements of sparseness [53], [55] or independence [54] to a linear transformation. It is unlikely that biological vision would select linear filters for the early stages of processing if linearity were, by itself, an undesirable property.

The third reason for our belief that linearity is not a major limitation is the problem of invariance. When the feature transform T is a multiresolution 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 lowpass 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 Fig. 3.

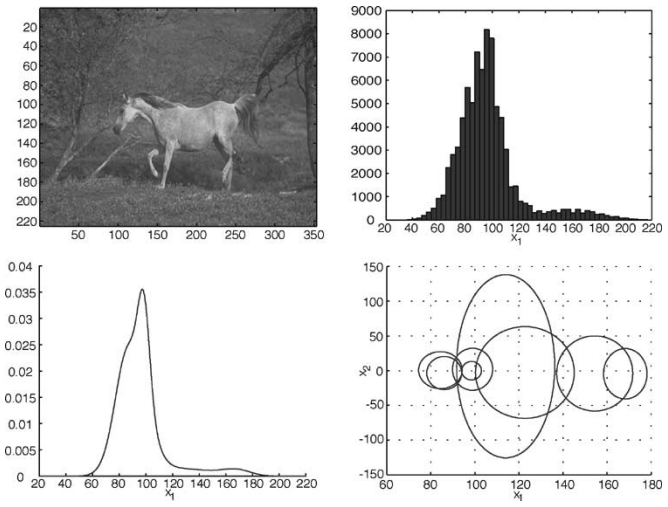

Fig. 3. (Top left) Image from the Corel database, (top right) its histogram, (bottom left) projection of the corresponding 64-dimensional embedded mixture onto the DC subspace, and (bottom right) projection onto the subspace of the two lower frequency coefficients. The embedded mixture describes the probability density of the discrete cosine transform coefficients derived from a collection of 8×8 blocks extracted from the image.

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

This observation suggests that a natural ordering for the subspaces generated by a multiresolution decomposition is by increasing frequency of the basis functions associated with those subspaces. The resulting *embedded multiresolution mixture* (EMM) model (embedded mixtures on a multiresolution feature space) is a generalization of the color histogram, where the additional dimensions capture the spatial dependencies that are crucial for fine image discrimination. Fig. 4 illustrates this point by presenting two images that have the exact same color histogram but are perceptually quite distinct. The advantage of the EMM generalization is that it 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 EMM representation is also invariant to all these transformations. However, by including high-frequency coefficients, it is possible to trade off invariance for Bayes error. Under this interpretation, the search for the MPE subspace of the previous section is a search for the subspace that achieves the optimal balance between the level of image detail required for recognition and that which starts to compromise invariance. There is, however, one slight limitation to the ordering by frequency, namely, that it is not straightforward to determine it for bases that are learned from training data. One well-known approximation that is popular in the retrieval literature [14], [20], [56] is to order the features by decreasing feature variance. Preliminary experiments, with transformations where the two orderings are easily determined, revealed no significant loss associated with this approximation, and we, therefore, adopt it in our implementation.

G. Multiresolution Feature Transforms

In this section, we briefly review the multiresolution feature transformations considered in the experimental section.

Definition 5: The discrete cosine transform (DCT) [57] of size n is the orthogonal transform whose basis functions are defined by

$$A(i,j) = \alpha(i)\alpha(j)\cos\frac{(2x+1)i\pi}{2n}\cos\frac{(2y+1)j\pi}{2n} \qquad 0 \le i,j,x,y < n \quad (29)$$

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 [27]. It is also possible to show that for certain classes of stochastic processes, the DCT converges asymptotically to the following transform [57].

Definition 6: Principal components analysis (PCA) is the orthogonal transform defined by

$$\mathbf{T} = \mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T \quad (30)$$

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 [42], [58].

Definition 7: A wavelet transform (WT) [59] is the orthogonal transform whose basis functions are defined by

$$A(i,j) = \sqrt{2^{k+l}}\Psi(2^k x - i)\Psi(2^l y - j)_{(0,0)\le(i,j)<(2^k,2^l)}^{0\le k,l<\log_2 n} \quad (31)$$

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 [53]. Eliminating such dependencies is the goal of independent component analysis (ICA).

Definition 8: ICA [60] is a feature transform such that

$$P_{\mathbf{X}}(\mathbf{x}) = \prod_i P_{\mathbf{X}_i}(\mathbf{x}_i) \quad (32)$$

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 that is guaranteed to be zero if the inputs are independent. Examples of

such contrast functions are higher order correlations and information-theoretic objective functions [60]. In this work, we consider representatives from the two types: the method developed by Comon [61], which uses a contrast function based on high-order cumulants, and the FastICA algorithm [62], which relies on the negative entropy of the features.

V. Experimental Evaluation

In this section, we present an experimental evaluation of the MPE retrieval architecture and compare it against various empirical retrieval techniques in common use.

A. Experimental Setup

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 an empirical technique that is commonly used in each retrieval domain and compared its performance with that of MPE retrieval. The implementation of the latter was as follows.

All images were normalized to the sizes 240×360 or 360×240. On databases containing color images, these were converted from the original RGB to the YBR color space. The image observations were 8×8 patches obtained with a sliding window moved by two pixels in a raster scan fashion (with a vertical interval of two lines), leading to a sample of about 20 000 observations per image. A PCA was applied to each color channel, the resulting 64 features ordered by decreasing variance, features from the different channels interleaved according to the pattern YBRYBR..., and the first 64 features of this pattern were kept, resulting in a 64-dimensional feature space (all 64 features were kept on Brodatz, where all images are grayscale). Mixtures of eight Gaussians with diagonal covariance were learned for all images with the EM algorithm [63] initialized with the generalized Loyd algorithm [64] according to the codeword splitting procedure discussed in [65]. After learning the initial set of means, all the vectors in the training set were assigned to the closest (in the Euclidean sense) mean vector, the sample covariances resulting from this assignment were used as initial estimate for the covariances, and the relative frequencies of the assignments were used as initial estimates for the mixture probabilities. Each image in the retrieval database was considered as a different class, leading to $M = 784$ classes on Brodatz, $M = 1200$ on Corel, and $M = 900$ on Columbia. Note that these are large-scale classification problems that simply cannot be handled by many of the existing supervised learning or feature selection techniques.

To evaluate the retrieval performance, we relied on standard precision/recall curves. In all the databases considered, there is clear ground truth regarding which images are relevant to a given query (e.g., images labeled as belonging to the same concept on Corel or different views of the same object on Columbia), and we used it to measure precision and recall. Each database was split into a training and test set, the images in the test set serving as queries for performance evaluation. We refer to this set as the *query database*. For the cross-validation procedure required by FS, the training set was further split into a training set, the *retrieval database*, and a set of cross-validation images (the *cross-validation database*). This split was performed in a manner similar to that of n-fold cross-validation [7], but to limit the computation, we did not iterate over different cross-validation sets. The whole process can be summarized as follows. For training, we considered only the retrieval and cross-validation databases, and the procedures of Section IV-E were used to determine the MPE feature subspace. The query and retrieval databases were then used to compare the performance of MPE retrieval with that of the competing techniques selected from the literature.

The specific organization of the databases and the empirical techniques against which MPE retrieval was compared for each database were as follows. The 1008 images in Brodatz were divided into a query database of 112, a cross-validation database of 112, and a retrieval database of 784 images. Various previous studies have identified the combination of 1) the coefficients of the least squares fit of an MRSAR model to each texture and 2) the Mahalanobis distance, as a top performer in this database [18], [22], [41]. We followed closely the implementation of [18], [22], [41], but preliminary experiments revealed that the performance of the MRSAR is significantly dependent on the modeling of the full covariance matrix. Hence, we have used full covariance matrices in all cases where the feature set was MRSAR.

The Columbia database was also split into three subsets: a query and a cross-validation database containing a single view of each of the 100 objects available and a retrieval database containing nine views (separated by 40°) of each object. It was chosen because it is a database where the histogram-based methods that are very popular in the retrieval literature [4] tend to perform well, allowing a comparison of MPE retrieval against these techniques. For color histogramming, the three-dimensional 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 the widely used histogram intersection (HI) metric, following the implementation of [13].

From Corel we selected 15 concepts[3] leading to a total of 1 500 images. Of these, 10% were used on the query database and 10% on the cross-validation, leaving the remaining 80% for retrieval database. In addition to the texture and color-based approaches, this database allowed the comparison of MPE retrieval against a popular empirical approach that jointly models the two attributes: the color correlogram proposed in [66], whose implementation was replicated in the experiments below.

[3]"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."

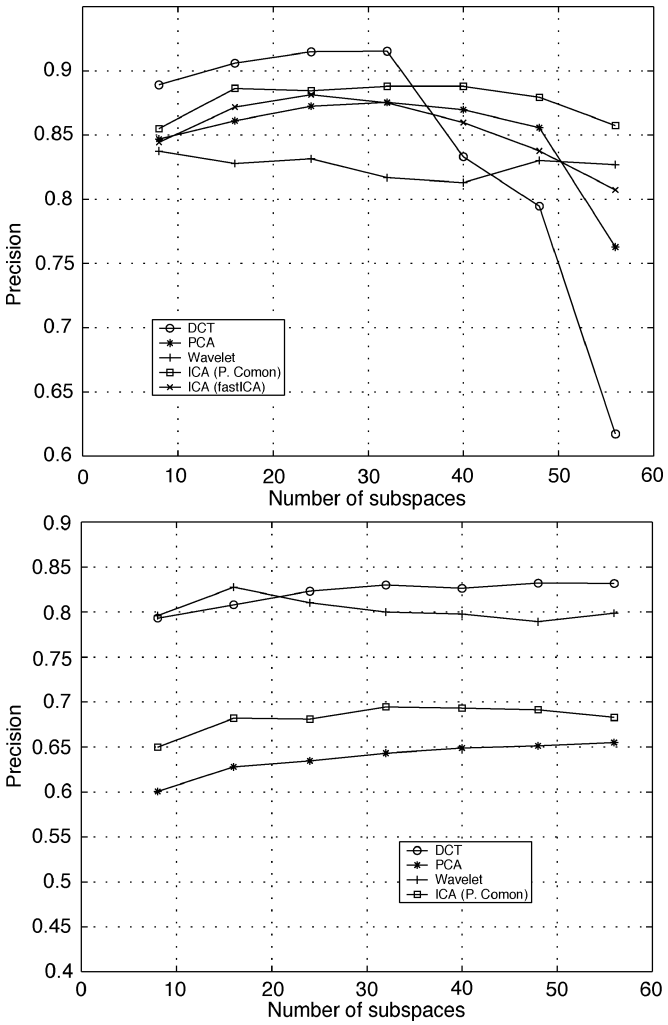

Fig. 5. (Top) Precision, at 30% recall on Brodatz. (Bottom) Precision, at 10% recall, on Corel.

B. Feature Transformation

We start with a set of results that illustrate 1) the importance of relying on a diverse dictionary of feature transforms as a means to achieve high retrieval accuracy over a diverse set of databases and 2) how the performance of a given transform can vary significantly with both the type of database and the selected subspace dimension. These results were obtained during the cross-validation stage, i.e., using the images in the cross-validation database as queries. For each query, we measured precision at various levels of recall. The precision/recall (PR) curves were then averaged over all queries to generate an average PR curve for each feature transform. Fig. 5 presents the curves of precision, as a function of subspace dimension, at 30% recall on Brodatz and 10% recall on Corel (the relative precision values obtained with the various transformations did not vary significantly with the level of recall).

The precision curves comply with the theoretical arguments of Section III-B1. 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 for all transformations (there is a large increase in precision from one to eight dimensions on both cases that we omit for clarity of the graph). Other than this, there are two other interesting observations. The first is that for a given database, a poor choice of transformation can lead to significant degradation of retrieval performance. For example, the peak precision of the worst transformation (wavelet) on Brodatz is 10% below that of the best (DCT), and on Corel, the variation is almost 20%. Furthermore, while the wavelet basis has the worst performance on Brodatz, it is one of the top two feature sets on Corel. On the other hand, ICA does better on Brodatz than on Corel. These are drastic variations in retrieval accuracy, which would be difficult to anticipate in the absence of this cross-validation stage. Second, even for a given feature 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. Overall, these observations emphasize the importance of relying on an optimal feature selection algorithm (under an optimality criterion that is sensible for retrieval) when the goal is to design robust retrieval systems applicable to a large range of image databases.

C. Comparison to Standard Solutions

In this section, we compare the performance of the MPE retrieval architecture proposed in this work with those of two empirical techniques discussed above (MRSAR and HI) in the specific databases where the latter work best: texture (Brodatz) for MRSAR and color (Columbia) for HI. Fig. 6 presents the resulting PR curves, showing that MPE retrieval achieves equivalent performance or actually outperforms the best of the two other approaches in each image domain. This indicates that the MPE architecture performs well for both color and texture and should therefore do well on a large spectrum of databases. Visual inspection of the retrieval results suggests that, also along the dimension of perceptual relevance, MPE retrieval clearly outperforms the MRSAR and histogram-based approaches. Fig. 7 presents representative examples of the three of major advantages of the MPE retrieval system:

1) When it makes errors, these tend to be perceptually less disturbing 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.
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.

The two pictures on the left column exemplify how MPE retrieval can lead to perceptually pleasing retrieval results, even when the PR performance is only mediocre. In this case, while HI retrieves several objects unrelated to the query, MPE 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, the embedded multiresolution mixture representation is able to capture the local appearance of the object surface. Hence, it tends to match surfaces with the same shape, texture, and reflection properties. This is not possible with color histograms.

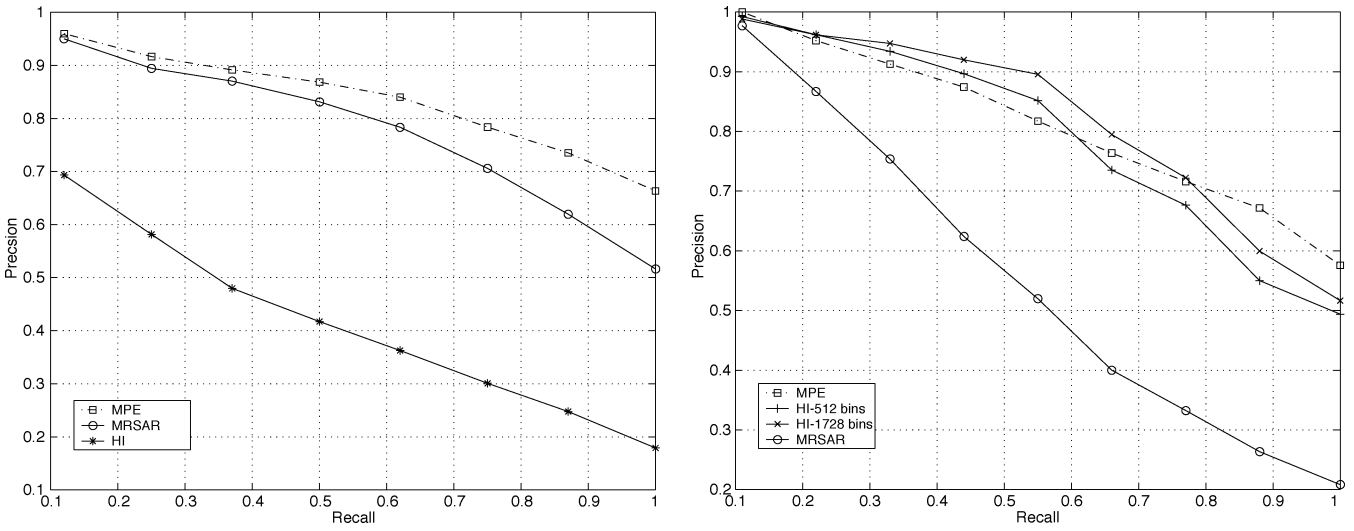

Fig. 6. PR measured for the MPE, MRSAR, and HI retrieval architectures. (Left) Curves from Brodatz, where the best results for HI (which are shown) were obtained with histograms of 192 bins. The features selected by MPE were the DCT set with 32 subspaces. (Right) Curves from Columbia where best HI results were obtained with histograms of 1728 bins. There was, however, a wide range of the number of bins where the performance was nearly constant, as illustrated by the second curve, which was obtained with histograms of 512 bins. The features selected by MPE were the DCT set with 48 subspaces.

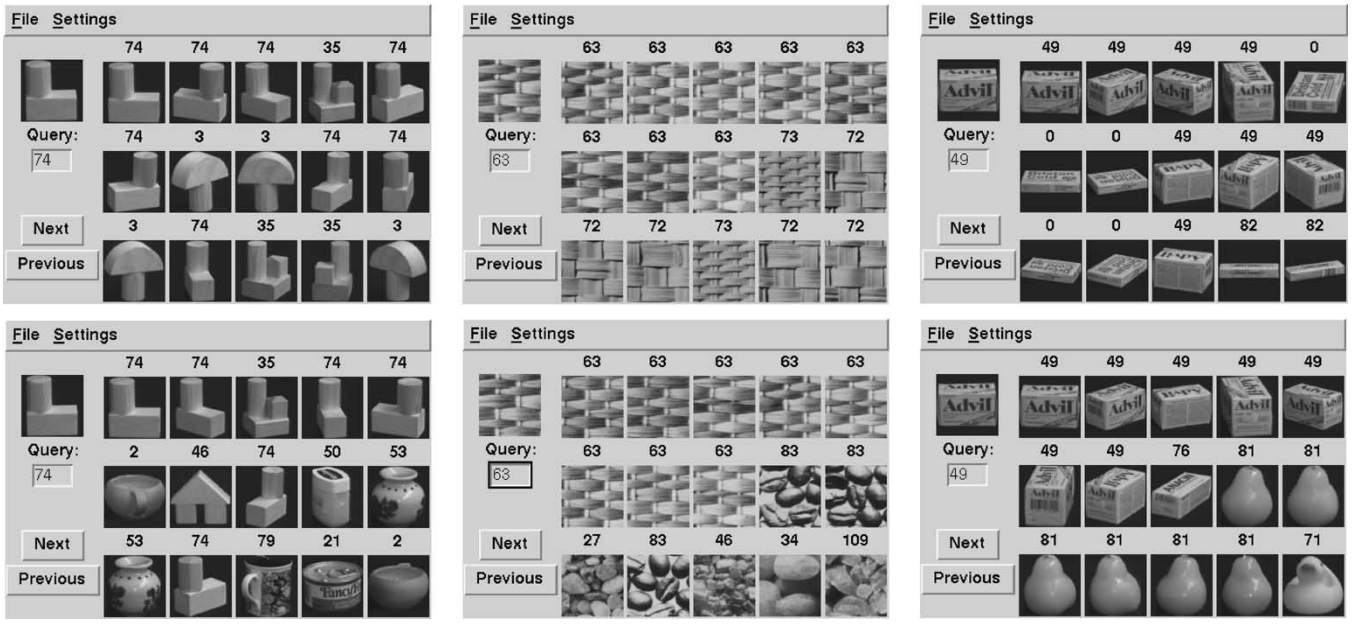

Fig. 7. (Top) Comparison of MPE retrieval results with (bottom) those of HI on Columbia and MRSAR on Brodatz.

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 ranks all the images in the query class at the top but produces poor matches after that. On the other hand, MPE 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. On the other hand, because a good match under MPE retrieval implies that the query and retrieved images should populate the space of spatial frequencies in a similar fashion, this approach tends to group images that have energy along the same orientations and a frequency spectrum with the same types of periodicities. These characteristics are known to be relevant for human judgments of similarity [22].

Finally, the pictures on the right column illustrate how, even when it has higher PR, HI can lead to perceptually poorer results than the MPE approach. 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, MPE retrieval only retrieves boxes, although not in the best possible order.

D. Generic Retrieval Solutions

We finalize with results from the Corel database. Fig. 8 presents a comparison, in terms of PR, of MRSAR, HI, the color correlogram, and MPE retrieval. It is clear that the texture model alone performs very poorly, color histogramming does

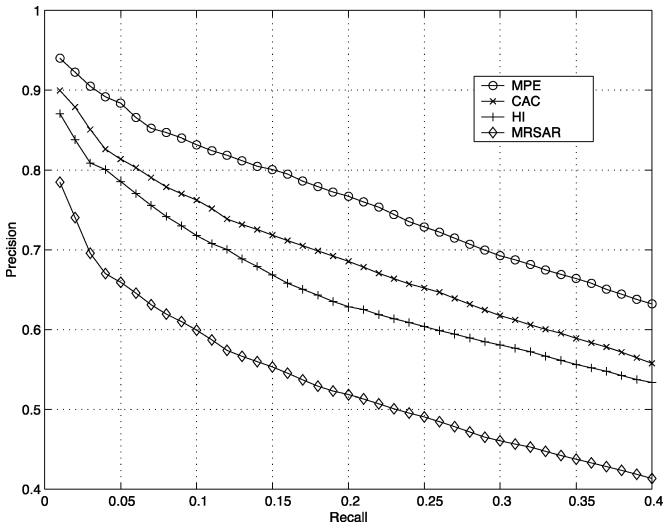

Fig. 8. PR on Corel for MRSAR, HI (512 bin histograms), color correlogram (CAC), and MPE retrieval. The features selected by MPE were the DCT set with 46 subspaces.

significantly better, and the correlogram further improves performance by about 5%. However, all the empirical approaches are significantly less effective than MPE retrieval.

APPENDIX

A. Proof of Theorem 4

Proof: The theorem follows from the application of two bounds. The first is that for a problem with class conditional densities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, equiprobable classes $P_Y(i) = 1/M$, $\forall i$, class-conditional density estimates $\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)$, and a feature space $\mathcal{X}$

$$P_{\mathbf{X},Y}\left(g(\mathbf{X}) \neq Y\right) - L^*_{\mathcal{X}} \leq \frac{1}{M}\sum_i \int \left|P_{\mathbf{X}|Y}(\mathbf{x}|i) - \hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)\right| d\mathbf{x}.$$

See [12] for the case where $M = 2$ and [27] for an extension to multiple classes. The second is a well-known bound in information theory, which is usually referred to as Pinsker's inequality; see, e.g., [28, Lemma 12.6.1] or [30, Th. 7.11.1]

$$\int |P_{\mathbf{X}}(\mathbf{x}) - Q_{\mathbf{X}}(\mathbf{x})|\, d\mathbf{x} \leq \sqrt{2\ln 2 KL\left[P_{\mathbf{X}}(\mathbf{x})||Q_{\mathbf{X}}(\mathbf{x})\right]}.$$

■

B. Proof of Theorem 5

Proof: The fact that the sequence of vector spaces is embedded follows from (12) since, $\forall i \in \{1, \ldots, d-1\}$

$$\mathcal{X}_i = \pi_i^{i+1}(\mathcal{X}_{i+1}) \tag{33}$$

and consequently, there is a sequence of one-to-one mappings

$$\epsilon_i(\mathbf{x}) = (\mathbf{x}, 0) \tag{34}$$

for which

$$\epsilon_i(\mathcal{X}_i) \subset \mathcal{X}_{i+1}. \tag{35}$$

Inequality (13) then follows from (33), (6), and the fact that the mappings $\pi_i^{i+1}(\mathbf{x})$ are noninvertible. To prove (16), we start from Theorem 4, i.e.,

$$\Delta_{g_i,\mathcal{X}_i} = \frac{\sqrt{2\ln 2}}{M}\sum_k \sqrt{KL\left[P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)||\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)\right]} \tag{36}$$

where $P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)$ is the class-conditional likelihood function for $\mathbf{X}_i$ under class k. Since, from (33), $\mathbf{X}_{i+1} = (\mathbf{X}_i, X_{i+1})$, where X_{i+1} is the $i+1$th coordinate of $\mathbf{X}_{i+1}$, we have

$$\begin{aligned}
&\mathrm{KL}\left[P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)||\hat{p}_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)\right] \\
&= \int P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)\log\frac{P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)}{\hat{p}_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)}d\mathbf{x}_{i+1} \\
&= \int P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)\log\frac{P_{X_{i+1}|\mathbf{X}_i,Y}(x_{i+1}|\mathbf{x}_i,k)}{\hat{p}_{X_{i+1}|\mathbf{X}_i,Y}(x_{i+1}|\mathbf{x}_i,k)}dx_{i+1}d\mathbf{x}_i \\
&\quad + \int P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)\log\frac{P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)}{\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)}dx_{i+1}d\mathbf{x}_i \\
&= \int P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k) \\
&\qquad \times \log\frac{P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)}{\hat{p}_{X_{i+1}|\mathbf{X}_i,Y}(x_{i+1}|\mathbf{x}_i,k)P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)}dx_{i+1}d\mathbf{x}_i \\
&\quad + \int P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)\log\frac{P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)}{\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)}d\mathbf{x}_i \\
&= \mathrm{KL}\left[P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)|| \right. \\
&\qquad \left. \hat{p}_{X_{i+1}|\mathbf{X}_i,Y}(x_{i+1}|\mathbf{x}_i,k)P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)\right] \\
&\quad + \mathrm{KL}\left[P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)||\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)\right] \\
&\geq \mathrm{KL}\left[P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)||\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)\right]
\end{aligned}$$

where we have used the nonnegativity of the KL divergence [28]. It follows from the fact that the square root is a monotonically increasing function that

$$\sqrt{\mathrm{KL}\left[P_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)||\hat{p}_{\mathbf{X}_{i+1}|Y}(\mathbf{x}_{i+1}|k)\right]} \geq \sqrt{\mathrm{KL}\left[P_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)||\hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}_i|k)\right]}$$

which, combining with (36), leads to (16). ■

C. Proof of Lemma 1

Proof: From the properties of symmetric block matrices [67], it is known that if

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{bmatrix}$$

where $\mathbf{A}$ and $\mathbf{D}$ are symmetric matrices, then

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1}(\mathbf{I} + \mathbf{B}\mathbf{P}^{-1}\mathbf{B}^T\mathbf{A}^{-1}) & -\mathbf{A}^{-1}\mathbf{B}\mathbf{P}^{-1} \\ -\mathbf{P}^{-1}\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{P}^{-1} \end{bmatrix} \tag{37}$$

$$= \Gamma(\mathbf{A}^{-1}) + \mathbf{E}\mathbf{P}^{-1}\mathbf{E}^T \tag{38}$$

and $|\mathbf{M}| = |\mathbf{A}||\mathbf{P}|$ with

$$\Gamma(\mathbf{A}^{-1}) = \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} \mathbf{A}^{-1}\mathbf{B} \\ -\mathbf{I} \end{bmatrix}$$

and $\mathbf{P} = \mathbf{D} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}$. Hence, for any vector $\mathbf{z}^T = [\mathbf{x}^T\mathbf{y}^T]$, where $\mathbf{x}$ and $\mathbf{y}$ have the appropriate lengths for $||\mathbf{z}||_{\mathbf{M}}$ to make sense

$$\begin{aligned}||\mathbf{z}||_{\mathbf{M}} &= ||\mathbf{x}||_{\mathbf{A}} + (\mathbf{B}^T\mathbf{A}^{-1}\mathbf{x} - \mathbf{y})^T\mathbf{P}^{-1}(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{x} - \mathbf{y}) \\ &= ||\mathbf{x}||_{\mathbf{A}} + ||\mathbf{B}^T\mathbf{A}^{-1}\mathbf{x} - \mathbf{y}||_{\mathbf{P}}. \end{aligned} \quad (39)$$

Using the decomposition

$$\mathbf{\Pi}_j = \begin{bmatrix} \mathbf{\Pi}_{j-1} \\ \mathbf{e}_j^T \end{bmatrix}$$

where $\mathbf{e}_j$ is the jth vector of the canonical basis of $\mathbb{R}^d$ (jth coordinate equal to 1, all others to 0), and defining $\mathbf{S}_j = \mathbf{\Pi}_j\mathbf{\Sigma}\mathbf{\Pi}_j^T$, it follows that

$$\mathbf{S}_j = \begin{bmatrix} \mathbf{S}_{j-1} & \mathbf{u}_{j-1} \\ \mathbf{u}_{j-1}^T & \sigma_{j,j} \end{bmatrix} \quad \text{and} \quad \mathbf{\Pi}_j\mathbf{d} = \begin{bmatrix} \mathbf{\Pi}_{j-1}\mathbf{d} \\ d_j \end{bmatrix}$$

where d_j is the jth element of $\mathbf{d}$. Making $\mathbf{M} = \mathbf{S}_j$, $\mathbf{A} = \mathbf{S}_{j-1}$, $\mathbf{B} = \mathbf{u}_{j-1}$, $\mathbf{D} = \sigma_{j,j}$, and defining $p_j = \mathbf{P}$, and $\boldsymbol{\psi}_j = \mathbf{E}$, it follows that

$$\begin{aligned} \boldsymbol{\psi}_j^T &= \left(\mathbf{u}_{j-1}^T\mathbf{S}_{j-1}^{-1}, -1\right) \\ p_j &= \sigma_{j,j} - ||\mathbf{u}_{j-1}||_{\mathbf{S}_{j-1}} \\ &= -\left(\mathbf{u}_{j-1}^T, \sigma_{j,j}\right)\boldsymbol{\psi}_j \\ \mathbf{S}_j^{-1} &= \Gamma\left(\mathbf{S}_{j-1}^{-1}\right) + \frac{1}{p_j}\boldsymbol{\psi}_j\boldsymbol{\psi}_j^T. \end{aligned}$$

Letting $\mathbf{z} = \mathbf{\Pi}_j\mathbf{d}$, $\mathbf{x} = \mathbf{\Pi}_{j-1}\mathbf{d}$, $\mathbf{y} = d_j$, and applying (39)

$$||\mathbf{\Pi}_j\mathbf{d}||_{\mathbf{S}_j} = ||\mathbf{\Pi}_{j-1}\mathbf{d}||_{\mathbf{S}_{j-1}} + \frac{1}{p_j}\left(\boldsymbol{\psi}_j^T\mathbf{\Pi}_j\mathbf{d}\right)^2.$$

Since $\mathcal{M}_j = ||\mathbf{\Pi}_j\mathbf{d}||_{\mathbf{S}_j}$, this leads to (23)–(26). Furthermore, from $|\mathbf{M}| = |\mathbf{A}||\mathbf{P}|$, it follows that $|\mathbf{S}_j| = p_j|\mathbf{S}_{j-1}|$, which leads to (27). Finally, since the steps of (23)–(27) have complexity $O(j)$ or $O(j^2)$, the overall complexity is $O(\sum_{j=1}^{d} j^2) = O(d(d+1)(2d+1)/6 = O(d^3)$ ■

D. Proof of Theorem 6

Proof: Since all transformations in $\mathcal{T}$ are invertible, for any pair $(T^{(l)}, T^{(m)})$, where $T^{(l)} : \mathcal{Z} \to \mathcal{X}^{(l)}$ and $T^{(m)} : \mathcal{Z} \to \mathcal{X}^{(m)}$, the transformation $T^{(l,m)} = T^{(m)} \circ (T^{(l)})^{-1}$ maps $\mathcal{X}^{(l)}$ into $\mathcal{X}^{(m)}$. It follows from Property 1 that if $\mathbf{X}^{(l)} \in \mathcal{X}^{(l)}$ is distributed according to a Gauss mixture with parameters $\{\lambda_i, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}$, then $\mathbf{X}^{(m)} = T^{(l,m)}(\mathbf{X}^{(l)})$ is distributed according to a Gauss mixture with parameters $\{\lambda_i, \mathbf{T}^{(l,m)}\boldsymbol{\mu}_i, \mathbf{T}^{(l,m)}\mathbf{\Sigma}_i(\mathbf{T}^{(l,m)})^T\}$. From Theorem 5, the sequence $\mathcal{X}_j^{(m)} = \pi_j^d(\mathcal{X}^{(m)})$ is a sequence of embedded spaces, and (28) follows from Property 1. The recursion for the computation of $P_{\mathbf{X}_j^{(m)}}(\mathbf{\Pi}_j\mathbf{T}^{(l,m)}\mathbf{x})$, $j = 1, \ldots, d$ follows directly from Lemma 1 by making the change of variables $\mathbf{x} = \mathbf{T}^{(l,m)}\mathbf{x}$, $\boldsymbol{\mu} = \mathbf{T}^{(l,m)}\boldsymbol{\mu}$, and $\mathbf{\Sigma} = \mathbf{T}^{(l,m)}\mathbf{\Sigma}(\mathbf{T}^{(l,m)})^T$. Finally, since for each m the cost of the recursion is $O(d^3)$ as well as for the F basis, the overall cost is $O(Fd^3)$. ■

ACKNOWLEDGMENT

The author would like to acknowledge various insights and criticism provided by A. Lippman, R. Gray, A. Bobick, M. Kunt, and G. Carneiro. G. Carneiro also provided invaluable help in many of the experiments reported in the paper. Finally, three anonymous reviewers provided various comments and pointers to the literature that significantly strengthened the quality of the presentation.

REFERENCES

[1] 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 *Proc. SPIE Storage Retrieval Image Video Databases*, San Jose, CA, 1993, pp. 173–181.

[2] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: content-based manipulation of image databases," *Int. J. Comput. Vision*, vol. 18, no. 3, pp. 233–254, June 1996.

[3] R. Picard, "Light-years from Lena: Video and image libraries of the future," in *Proc. Int. Conf. Image Processing*, Washington, DC, Oct. 1995.

[4] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval: the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1349–1380, Dec. 2000.

[5] N. Vasconcelos and M. Kunt, "Content-based retrieval from image databases: Current solutions and future directions," in *Proc. Int. Conf. Image Process.*, Thessaloniki, Greece, 2001.

[6] T. Huang and X. Zhou, "Image retrieval and relevance feedback: from heuristic weight adjustment to optimal learning methods," in *Proc. IEEE Int. Conf. Image Process.*, Thessaloniki, Greece, 2001.

[7] G. Johh, R. Kohavi, and K. Phleger, "Irrelevant features and the feature subset problem," in *Proc. Int. Conf. Machine Learning*, Bari, Italy, 1994.

[8] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intell.*, vol. 97, pp. 245–272, 1997.

[9] V. A. Kotel'nikov, *The Theory of Optimum Noise Immunity*. New York: McGraw-Hill, 1959.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.

[11] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2001.

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

[13] M. Swain and D. Ballard, "Color indexing," *Int. J. Comput. Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[14] J. De Bonet and P. Viola, "Structure driven image database retrieval," in *Proc. Neural Inform. Process. Syst.*, vol. 10, Denver, CO, 1997.

[15] T. Gevers and A. Smeulders, "PickToSeek: combining color and shape invariant features for image retrieval," *IEEE Trans. Image Processing*, vol. 9, pp. 102–119, Jan. 2000.

[16] J. Hafner, H. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 729–736, July 1995.

[17] A. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognit. J.*, vol. 29, pp. 1233–1244, August 1996.

[18] W. Ma and H. Zhang, "Benchmarking of image features for content-based retrieval," in *Proc. 32nd Asilomar Conf. Signals, Syst., Comput.*, Asilomar, CA, 1998.

[19] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," in *Proc. Int. Conf. Computer Vision*, Korfu, Greece, 1999, pp. 1165–1173.

[20] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 644–655, Sept. 1998.

[21] J. Smith and S. Chang, "VisualSEEk: A fully automated content-based image query system," in *ACM Multimedia*, Boston, MA, 1996, pp. 87–98.

[22] F. Liu and R. Picard, "Periodicity, directionality, and randomness: wold features for image modeling and retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 722–733, July 1996.

[23] B. Manjunath and W. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 837–842, Aug. 1996.

[24] N. Vasconcelos, "Image indexing with mixture hierarchies," in *Proc. IEEE Comput. Vision Pattern Recogn. Conf.*, Kawai, HI, 2001.

[25] N. Vasconcelos and A. Lippman, "Learning over multiple temporal scales in image databases," in *Proc. Eur. Conf. Comput. Vision*, Dublin, Ireland, 2000.

[26] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliifs, NJ: Prentice-Hall, 1993.
[27] N. Vasconcelos, "Bayesian models for visual information retrieval," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 2000.
[28] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[29] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. CT-15, pp. 52–60, Feb. 1967.
[30] R. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison Wesley, 1991.
[31] J. Simonoff, *Smoothing Methods in Statistics*. New York: Springer-Verlag, 1996.
[32] D. Scott, *Multivariate Density Estimation*. New York: Wiley-Interscience, 1992.
[33] Q. Li, "Estimation of mixture models," Ph.D. dissertation, Yale University, New Haven, CT, 1999.
[34] V. Guillemin, *Differential Topology*. Harlow, U.K.: Pearson Education, 1974.
[35] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[36] G. Lugosi, "Pattern classification and learning theory," in *Principles of Nonparametric Learning*, L. Gyorfi, Ed. New York: Springer, 2002, ch. 1.
[37] D. Titterington, A. Smith, and U. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.
[38] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 153–158, Feb. 1997.
[39] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image classification for content-based indexing," *IEEE Trans. Image Processing*, vol. 10, pp. 117–130, Jan. 2001.
[40] T. Cover and J. Van Campenhout, "On the possible orderings in the measurement selection problem," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, pp. 657–661, Sept. 1977.
[41] R. Picard, T. Kabir, and F. Liu, "Real-time recognition with the entire Brodatz texture database," in *Proc. IEEE Conf. Comput. Vision*, New York, 1993.
[42] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neurosci.*, vol. 3, 1991.
[43] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition*, San Juan, Puerto Rico, 1997.
[44] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. London, U.K.: Chapman & Hall, 1993.
[45] J. Mao and A. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recogn.*, vol. 25, no. 2, pp. 173–188, 1992.
[46] D. Hubel and T. Wiesel, "Brain mechanisms of vision," *Sci. Amer.*, Sept. 1979.
[47] D. Sagi, "The psychophysics of texture segmentation," in *Early Vision and Beyond*, T. Papathomas, Ed. Cambridge, MA: MIT Press, 1996, ch. 7.
[48] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *J. Opt. Soc. Amer.*, vol. 7, no. 5, pp. 923–932, May 1990.
[49] J. Bergen and E. Adelson, "Early vision and texture perception," *Nature*, vol. 333, no. 6171, pp. 363–364, 1988.
[50] I. Fogel and D. Sagi, "Gabor filters as texture discriminators," *Biol. Cybern.*, vol. 61, pp. 103–113, 1989.
[51] A. Sutter, J. Beck, and N. Graham, "Contrast and spatial variables in texture segregation: testing a simple spatial-frequency channels model," *Perceptual Psychophys.*, vol. 46, pp. 312–332, 1989.
[52] J. Bergen and M. Landy, "Computational modeling of visual texture segregation," in *Computational Models of Visual Processing*, M. Landy and J. Movshon, Eds. Cambridge, MA: MIT Press, 1991.
[53] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
[54] A. Bell and T. Sejnowski, "The independent components of natural scenes are edge filters," *Vision Res.*, vol. 37, no. 23, pp. 3327–3328, Dec. 1997.
[55] D. Field, "What is the goal of sensory coding?," *Neural Comput.*, vol. 6, no. 4, pp. 559–601, Jan. 1989.
[56] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: color- and texture-based image segmentation using EM and its application to image querying and classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 1026–1038, Aug. 2002.
[57] N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[58] H. Murase and S. Nayar, "Visual learning and recognition of 3-D objects from appearance," *Int. J. Comput. Vision*, vol. 14, pp. 5–24, 1995.
[59] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, July 1989.
[60] J. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 90, pp. 2009–2026, Oct. 1998.
[61] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.
[62] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.
[63] A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. B-39, 1977.
[64] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992.
[65] R. Gray, "Vector quantization," *IEEE Signal Processing Mag.*, vol. 1, pp. 4–29, Apr. 1984.
[66] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Spatial color indexing and applications," *Int. J. Comput. Vision*, vol. 35, no. 3, pp. 245–268, Dec. 1999.
[67] M. Artin, *Algebra*. Englewood Cliffs, NJ: Prentice-Hall, 1991.

Nuno Vasconcelos (M'00) received the licenciatura degree in electrical engineering and computer science from the Universidade do Porto, Porto, Portugal, in 1988 and the M.Sc. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1993 and 2000, respectively.

From 2000 to 2002, he was a member of the research staff at the Compaq Cambridge Research Laboratory, which in 2002 became the HP Cambridge Research Laboratory. In 2003, he joined the Electrical and Computer Engineering Department, the University of California, San Diego, La Jolla. He has worked in various areas including signal processing and compression, computer vision, machine learning, and multimedia. His current interests are in statistical signal processing, statistical computer vision, machine learning, large signal repositories, and multimedia.