On the Efficient Evaluation of Probabilistic Similarity Functions for Image Retrieval

Nuno Vasconcelos, *Member, IEEE*

Abstract—**Probabilistic approaches are a promising solution to the image retrieval problem that, when compared to standard retrieval methods, can lead to a significant gain in retrieval accuracy. However, this occurs at the cost of a significant increase in computational complexity. In fact, closed-form solutions for probabilistic retrieval are currently available only for simple probabilistic models such as the Gaussian or the histogram. We analyze the case of mixture densities and exploit the asymptotic equivalence between likelihood and Kullback–Leibler (KL) divergence to derive solutions for these models. In particular, 1) we show that the divergence can be computed exactly for vector quantizers (VQs) and 2) has an approximate solution for Gauss mixtures (GMs) that, in high-dimensional feature spaces, introduces no significant degradation of the resulting similarity judgments. In both cases, the new solutions have closed-form and computational complexity equivalent to that of standard retrieval approaches.**

Index Terms—**Bayes classifier, Gauss mixture (GM), image databases, Kullback–Leibler (KL) divergence, maximum $a\,posteriori$ probability (MAP) similarity, probabilistic image retrieval, vector quantizer (VQ).**

I. Introduction

DATABASE theory and the design of the various components that constitute a database system have been two enormously successful areas in computer science. However, because it is oriented to text-based data structures, existing database technology cannot fully address the challenges posed by modern databases that contain, in addition to text, multiple other signal modalities. Examples include multimedia databases containing audio, video, and graphics or life-sciences databases containing medical imagery and DNA information. While it is theoretically possible to simply annotate all the database content with text metadata and rely on traditional database architectures to later retrieve the desired information, in practice there are many situations in which such solutions are impossible or not cost effective. There are various reasons for this, including the enormous amounts of data involved (which make the annotation cost overwhelming) or the fact that various interpretations can be given to an image, a piece of music, or a DNA sample. Some of these interpretations can even be unknown at annotation time making it difficult, if not impossible, to predict the interpretation that a given user will have in mind at retrieval time. Due to these problems, an entirely new database search paradigm has been advocated by various researchers over the last decade, under the name of content-based retrieval [38], [42], [43], [49], [63]. The main idea is to augment the traditional text-based search paradigm with the ability to search by example, i.e., allowing users to express queries according to the similarity to user-provided examples. While we focus on the area of image retrieval, this type of search can be easily applied to other signal modalities, including audio [12], [31] or bio-informatics signals.

The design of an architecture for content-based image retrieval (CBIR) requires the specification of 1) an image representation suited for search and 2) a similarity function that, based on that representation, establishes a ranking of all images in the database according to their similarity to a set of query images. Since a natural goal for a retrieval system is to minimize the probability of retrieval error, the retrieval problem can be formally addressed with recourse to decision theory [60]. This implies a probabilistic formulation, where images are represented as observations from stochastic processes and the similarity function becomes the posterior probability of the query observations under the probabilistic models associated with the image classes in the database. Many observation spaces are possible, including the widely popular space of pixel colors inherent to the representation of an image by its color histogram [38], [49], [55], or various feature spaces designed to capture properties such as texture [30], [34], [35] or object shape [2], [23].

The appeal of decision-theoretic retrieval, to which we also refer to as probabilistic retrieval or maximum *a posteriori* probability (MAP) retrieval, derives from several properties of practical interest. In particular, it can be shown that 1) it does indeed minimize the probability of retrieval error, 2) it generalizes a significant number of other previously proposed retrieval approaches, 3) it establishes a common framework for handling global (based on entire images) and local (based on image regions) similarity, 4) it provides a natural foundation for the design of learning (relevance feedback) algorithms through belief propagation, and 5) it allows the natural integration of multiple content modalities (e.g., queries taking into account both images and the text of associated captions) [60]. There is, nevertheless, one significant hurdle to the practical implementation of probabilistic retrieval systems: the computational complexity of the MAP similarity function. Since the number of observations, Q, extracted from a query image can be very large (e.g., for color histogram methods each pixel originates an independent observation), the straightforward evaluation of the MAP similarity function has complexity $O(Q)$ per database class. This is usually overwhelming and, due to this limitation, probabilistic

Manuscript received March 17, 2003; revised February 11, 2004. The material in this paper was presented in part at the IEEE International Conference on Computer Vision, Vancouver, BC, Canada, 2001.

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

Communicated by E.-h. Yang, Guest Editor.

Digital Object Identifier 10.1109/TIT.2004.830760

similarity functions have not been widely used in the image retrieval literature, even though virtually all image representations in current use are probabilistic in nature [6], [17], [34], [37], [45], [47], [55].

Instead, the most popular strategy is to rely on a similarity function that takes as arguments the probabilistic models for both the query and database image class and produces a score derived from their parameters. Examples include the use of Euclidean distances between histograms (the most popular among which is the L^1 norm of the histogram difference commonly referred to in the literature as histogram intersection (HI)) [19], [55], the Mahalanobis distance between the mean value of the image features [34], [50], or metrics derived from related optimality criteria (e.g., the earth mover's distance between histograms introduced in [47]). Operating directly on models characterized by a small number P of parameters (e.g., histogram bins), these alternative metrics have the advantage of a reduced complexity $O(P)$, typically orders of magnitude smaller than $O(Q)$ (while millions of pixels may be extracted from an image, their color histogram typically has less than 256 bins). The cost is suboptimal performance in terms of probability of retrieval error. Indeed, it can be shown that many of these metrics can be derived from the MAP similarity function by making various assumptions and/or approximations (e.g., Gaussianity, linearizations, etc.) that are unrealistic for image data and/or discard information which is important for the evaluation of image similarity [60], [61]. Experimental evaluation has also confirmed that many of these alternative similarity functions cannot match the performance of the MAP criteria [45], [60], [64].

In this work, we seek ways to avoid the penalty in retrieval error inherent to the use of suboptimal similarity functions by deriving computationally efficient ways to evaluate the MAP function. The starting point is the well-known result that, up to a constant, the negative log-likelihood of the query observations under a given database image class converges asymptotically to the Kullback–Leibler (KL) divergence between the underlying query density and the density of that class. Since, for P-parameter models, one would expect the KL divergence to have computational cost $O(P)$, there is no *a priori* reason to believe that it should be more expensive than that of any other parametric solutions. However, the KL divergence between two probability density functions (pdfs) cannot always be expressed as a closed-form expression of their parameters. In fact, there is only a small set of models for which a closed-form expression is available. While this set includes models that have been widely applied to the CBIR problem, such as the Gaussian, the histogram, or variations/extensions [9], [20], [22], [34], such models have important limitations in the context of the CBIR problem. In particular, they either 1) are too simplistic to accurately describe the densities associated with real images (e.g., the Gaussian), 2) rely on assumptions that can severely compromise retrieval accuracy (e.g., independence between the components of the feature space) [62], or 3) are too rigid to be useful in the high-dimensional spaces required for accurate image discrimination (e.g., the histogram) [64].

We consider two models, vector quantizers (VQs) [14] and mixture densities [56], that overcome these fundamental limitations but for which no closed-form expression for the KL divergence is currently known. Both models can be seen as extensions of the histogram. However, unlike the histogram, they partition the feature space according to the distribution of the data, leading to density estimates whose complexity is determined by the complexity of this distribution (number of clusters that it contains) and not the dimension of the space itself. This enables estimates of reasonable accuracy on high-dimensional spaces, something that is impossible with histograms. The only, yet significant, difference between the two models is that while (like regular histograms) VQ-based density estimation procedures partition the space into mutually exclusive cells, mixture densities rely on soft partitions. We present two main results. The first is that a closed-form solution to the nearest neighbor problem, in the KL sense, does exist in the VQ case. Interestingly, this result is obtained by exploiting the relationships between the VQ and the Gauss mixture (GM) and therefore also provides new insights on the KL divergence between mixtures. This leads to the second result, a closed-form approximation for this quantity that is shown to be exact under two conditions. It is argued that these conditions are met approximately in high-dimensional spaces, where the retrieval performance of the new approximation is experimentally shown to be close to that of the MAP function. The practical consequence of these theoretical results is to enable probabilistic retrieval with sophisticated density models at a computational cost similar to that achieved with similarity functions in current use but higher retrieval accuracy.

The paper is organized as follows. Probabilistic retrieval is briefly reviewed in Section II, where we also introduce all the probabilistic models considered in this work and analyze their relationships. The closed-form solution for the nearest neighbor, in the KL sense, between VQs is derived in Section III, which also includes various intermediate results for the GM case. The asymptotic likelihood approximation (ALA) is then proposed for mixtures in Section IV, where the conditions under which it is exact are also derived. An experimental evaluation of the retrieval performance of the ALA is then presented on Section V where the latter is compared with the exact MAP solution and two other similarity functions commonly used in the literature. Finally, Section VI presents some conclusions and discusses further applications of this work.

II. Probabilistic Image Retrieval

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 n of them located in a predefined spatial neighborhood. We denote the space of observations by $\mathcal{Z} \subset \mathbb{R}^n$. The scalar n is always used to denote the dimension of the space $\mathcal{Z}$. Observations are mapped into feature vectors by a transformation $T : \mathcal{Z} \rightarrow \mathcal{X}$. We refer to $\mathcal{X}$ as the feature space, and $\boldsymbol{x} = T(\boldsymbol{z})$ a feature vector. Features are the elements of a feature vector. We associate a class indicator variable $Y \in \{1, \dots, M\}$ with the image classes in the database and denote the pdf of class i by $P_i(\boldsymbol{x})$. The pdf of the query feature vectors is denoted by $P(\boldsymbol{x})$. The following theorem is a well-known result from decision theory.

Theorem 1: Consider a feature space $\mathcal{X}$, a query set $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of N feature vectors drawn independently according to a pdf $P(\boldsymbol{x})$, a set of M database image classes with pdfs $P_i(\boldsymbol{x}), i = 1, \ldots, M$, and the set of similarity functions

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

Then, the similarity function that minimizes the probability of retrieval error, $P(g(\boldsymbol{X}) \neq Y)$, is the *Bayes* or MAP similarity function

$$\begin{aligned} g^*(\boldsymbol{X}) &= \arg\max_i P(Y = i|\boldsymbol{X}) \\ &= \arg\max_i \sum_k \log P_i(\boldsymbol{x}_k) + \log P(Y = i) \end{aligned} \quad (1)$$

Proof: See [8], [11], [13], among many other textbooks.

Retrieval systems based on the similarity function (1) are referred to as probabilistic, MAP retrieval systems. In the remainder of this paper, we will consider the classes to be *a priori* equally likely, in which case the prior probabilities $P(Y = i)$ can be ignored, but all results extend to the case of a nonuniform prior. Under the uniform assumption, the MAP similarity function is also referred to as the maximum-likelihood (ML) similarity function. We will also refer to $P(\boldsymbol{x})$ as the query density and to the density of the ith image class in the database $P_i(\boldsymbol{x})$ as the database density when the class label i is not relevant or can be inferred from context.

One immediate corollary of Theorem 1, that follows by straightforward application of the law of large numbers [10], is that ML similarity is asymptotically equivalent to maximizing the expected log (EL) of $P_i(\boldsymbol{x})$ under $P(\boldsymbol{x})$, i.e.,

$$g^*(\boldsymbol{X}) = \arg\max_i \mathrm{EL}\,(P\|P_i) = \arg\max_i \int P(\boldsymbol{x}) \log P_i(\boldsymbol{x}) d\boldsymbol{x}. \quad (2)$$

Because $P(\boldsymbol{x})$ is independent of i, this is the same as finding the database density that is closest to that of the query in the KL sense, i.e., the value of i that minimizes the KL divergence

$$\mathrm{KL}\,(P\|P_i) = \int P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{P_i(\boldsymbol{x})} d\boldsymbol{x}. \quad (3)$$

A. Applications of KL Divergence to Signal Matching

The application of the MAP and ML rules to the classification or matching of signals has a long history in communications and signal processing (see, e.g., [57]). The KL divergence and its applications to classification were introduced by Kullback in the context of the principle of minimum discrimination information (MDI) [25]. Given a density $P_1(\boldsymbol{x})$ and the set $\mathcal{M}$ of all densities that satisfy a constraint $\int T(\boldsymbol{x})P_2(\boldsymbol{x}) = \theta$, MDI seeks the density in $\mathcal{M}$ that is the "nearest neighbor" of $P_1(\boldsymbol{x})$ in the KL sense

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

Kullback showed that the minimum is 1) achieved by

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

where Z is a normalizing constant, $Z = \int e^{-\lambda T(\boldsymbol{x})} P_1(\boldsymbol{x}) d\boldsymbol{x}$, and λ a Lagrange multiplier [3] that weighs the importance of the constraint; and 2) equal to

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

Kupperman [25], [26] has shown that when all densities are members of the exponential family (a family that includes many of the common distributions of interest such as the Gaussian, Poisson, binomial, Rayleigh, and exponential among others [11]), MDI is equivalent to ML.

One of the earliest practical applications of this principle were in the areas of speech coding and recognition, where Itakura and Saito showed that the analysis step of linear predictive coding (LPC) is mathematically equivalent to a minimum distortion mapping that they referred to as the "error matching measure" [21], but is now commonly known as the "Itakura–Saito distortion." This similarity measure now forms the basis of most speech coding systems. Itakura and Saito derived the LPC analysis equations as an asymptotic approximation to the ML estimate of the parameters of a Gaussian autoregressive source with respect to a large training set of speech samples. Gray and colleagues then showed that the Itakura–Saito style of distortion measures are asymptotic MDI measures between the LPC model and sample autocorrelation of an observed speech frame, where the minimum is over the set of probability functions having the prescribed autocorrelation values [18].

All these developments assumed some form of Gaussianity, even though Gray's formulation did not require the original speech to be Gaussian, only its synthesized counterpart. More recently, the ML principle has been extensively used in areas such as speech recognition. In this context, the sequence of speech frames is modeled as a sequence of observations from a hidden Markov source, and the ML rule used for phoneme recognition [46]. Hidden Markov models are a generalization of the mixture models that we analyze in this work. Unlike the Gaussian case, the KL divergence between these models does not have a closed-form expression. Hence, while the asymptotic equivalence between maximizing likelihood and minimizing KL divergence still holds, it is difficult to exploit this connection in the design of efficient algorithms. This can be a major computational bottleneck, by making the complexity of the similarity function linear in the number of vectors to classify.

B. Probabilistic Models

While the complexity of (1) is linear in the cardinality N of the set of query vectors $\boldsymbol{X}$, the complexities of both (2) and (3) are only functions on the number of parameters in $P(\boldsymbol{x})$ and $P_i(\boldsymbol{x})$. Since the cardinality of the set of parameters is typically quite small, the maximization of EL (or minimization of KL divergence) is computationally more efficient whenever these quantities can be computed in closed form. The availability of closed-form solutions is, however, not universal, i.e., closed-form solutions only exist for some density families. In this subsection, we introduce the four types of probabilistic models considered in this work—the Gaussian, histogram, VQ, and GM—and analyze the relationships between them.

1) Mixture Densities: A mixture density [56] is a probabilistic model for stochastic processes with hidden structure. Associated with each image class[1] there is a hidden variable that defines a set of subclasses, which we denote by feature subclasses or clusters. The feature vectors that compose images from a given class are drawn in a sequence of two steps. First, one among the feature subclasses is selected according to a set of feature class probabilities. Feature vectors are then drawn according to a feature class-conditional density. Denoting the number of clusters by C, the sequence of feature class-conditional densities by $\{P(\boldsymbol{x}|\omega_c)\}_{c=1}^{C}$, and the feature class probabilities by $\{P(\omega_c)\}_{c=1}^{C}$, this can be expressed as

$$P(\boldsymbol{x}) = \sum_{c=1}^{C} P(\boldsymbol{x}|\omega_c)P(\omega_c). \tag{4}$$

The mixture model can account for nonstationary processes, e.g., an image containing various objects of different texture, and typically leads to a multimodal pdf. There is no closed-form solution to the EL or KL divergence between two mixture densities. We next show that the three other models under consideration are special cases of the mixture model.[2]

2) The Gaussian Model: By simply making $C = 1$, it is obvious from (4) that any parametric density is a particular case of the mixture model. In particular, we obtain a Gaussian of mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ when

$$P(\boldsymbol{x}|\omega_1) = \mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}||\boldsymbol{x}-\boldsymbol{\mu}||^2_{\boldsymbol{\Sigma}}} \tag{5}$$

where

$$||\boldsymbol{x} - \boldsymbol{\mu}||^2_{\Sigma} = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}). \tag{6}$$

The EL between two Gaussians is[3]

$$\begin{aligned}\mathrm{EL}\,(\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)) &= \\ -\frac{1}{2}\log(2\pi|\boldsymbol{\Sigma}_i|) - \frac{1}{2}\mathrm{trace}[\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}] &- \frac{1}{2}||\boldsymbol{\mu} - \boldsymbol{\mu}_i||^2_{\boldsymbol{\Sigma}_i}\end{aligned} \tag{7}$$

and a similar expression is available for the KL divergence [25].

3) Vector Quantizers (VQs): To analyze the relationship with VQ, we start by noticing that any mixture model induces a soft partition of the feature space. In particular, given a feature vector $\boldsymbol{x}$, the posterior probability assignment of that vector to each of the feature subclasses is

$$\begin{aligned}P(\omega_i|\boldsymbol{x}) &= \frac{P(\boldsymbol{x}|\omega_i)P(\omega_i)}{\sum_{k=1}^{C} P(\boldsymbol{x}|\omega_k)P(\omega_k)} \\ &= \begin{cases} \frac{1}{1+\sum_{k\neq i}\frac{P(\boldsymbol{x}|\omega_k)P(\omega_k)}{P(\boldsymbol{x}|\omega_i)P(\omega_i)}}, & \text{if } P(\boldsymbol{x}|\omega_i)P(\omega_i) > 0 \\ 0, & \text{otherwise.}\end{cases}\end{aligned} \tag{8}$$

This leads to an explicit connection with VQ.

Theorem 2: If $\boldsymbol{x}$ is a random vector distributed according to a GM

$$P_\epsilon(\boldsymbol{x}) = \sum_c P(\omega_c)\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c(\epsilon))$$

with covariance matrices

$$\boldsymbol{\Sigma}_c(\epsilon) = \epsilon\boldsymbol{I}, \qquad \forall c$$

then

$$\lim_{\epsilon\to 0} P_\epsilon(\omega_i|\boldsymbol{x}) = \begin{cases} 1, & \text{if } ||\boldsymbol{x} - \boldsymbol{\mu}_i||^2 \leq ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2 \forall k \neq i \\ 0, & \text{otherwise}\end{cases} \tag{9}$$

and

$$\lim_{\epsilon\to 0} P_\epsilon(\boldsymbol{x}) = \sum_{i=1}^{C} \delta(\boldsymbol{x} - \boldsymbol{\mu}_i)P(\omega_i). \tag{10}$$

where $\delta(\boldsymbol{x})$ is the Dirac delta function.

Proof: See Appendix I.

Equations (9) and (10) are a generative model for a VQ. In the VQ literature [14], (9) is known as the nearest neighbor condition (each point is assigned to the feature class associated with the nearest-neighbor codebook entry), and (10) as the centroid condition (this codebook entry is the mean of the cell associated with the feature class). The quantization operation consists of replacing each point by the codebook entry of the feature class to which it is assigned. The efficient evaluation of the EL and KL divergence between two VQs is the subject of Section III.

4) The Histogram Model: The histogram of a collection of feature vectors $\mathcal{S}$ is a vector $\boldsymbol{P} = \{p_1, \ldots, p_C\}$ associated with a partition of the feature space $\mathcal{X}$ into C cells, or bins, $\{\mathcal{X}_1, \ldots, \mathcal{X}_C\}$, where p_i is the percentage of vectors in $\mathcal{S}$ landing on cell $\mathcal{X}_i$. It provides an empirical estimate of the cluster probabilities $P(\omega_i)$ and, since all information other than these probabilities and the cell centroids $\boldsymbol{c}_i$ is discarded, has as underlying generative model

$$P(\boldsymbol{x}) = \sum_{i=1}^{C} \delta(\boldsymbol{x} - \boldsymbol{c}_i)P(\omega_i). \tag{11}$$

Comparing with (10), it is clear that the histogram is a particular case of the VQ model and, therefore, of the GM. In fact, the only difference with respect to the generic VQ model is that, in the histogram case, the cells into which the feature space is partitioned are defined arbitrarily and not learned from a training sample. While, conceptually, this is a small difference, in practice it can lead to substantially different retrieval performance. Because histogram partitions are arbitrary, there is no reason to have different ones for the different image classes. Hence, a universal partition (e.g., square cells of uniform size) is usually adopted for all classes. When this is the case, the KL divergence between two histograms is

$$\mathrm{KL}\,(P||P_i) = \sum_j P(\omega_j)\log\frac{P(\omega_j)}{P_i(\omega_j)}. \tag{12}$$

This solution, which we refer to as fixed partition or quantization, is clearly not ideal: for any given image class there will be

[1] For simplicity, since all results discussed in this section are class independent, we omit class subscripts from all pdfs.

[2] While some of these relationships are trivial or previously known, we include them for completeness.

[3] See Lemma 3 for a proof of a generalization of this result.

many empty cells and a few strongly populated ones. A more flexible approach, which is possible with VQ and mixture estimates, is to rely on an *adaptive partition* scheme where a different partition is learned for each image class. This is, in the context of this work, the true distinction between a VQ and an histogram. While VQ-based representations have been frequently used in the image retrieval literature [17], [22], [36], [45], [58], they typically rely on a fixed partition learned from a sample of all the image classes in database. This is only marginally different from using histograms and does not really address the major limitations of the histogram model. The problem is that it is not clear how to evaluate the KL divergence between two VQs or histograms defined on different partitions of the feature space.

C. Image Similarity Measures

Perhaps due to the lack of universal closed-form solutions, and despite its appeal as the decision function that minimizes the probability of retrieval error, the MAP similarity function has not received significant attention in the CBIR literature. An overwhelmingly more popular set of image similarity metrics is that of the L^p norms of the difference between densities

$$\mathcal{D}_p(P||P_i) = \arg\min_i \left(\int |P(\boldsymbol{x}) - P_i(\boldsymbol{x})|^p d\boldsymbol{x} \right)^{\frac{1}{p}}. \quad (13)$$

These norms are particularly common in the color-based retrieval literature, where they are used as metrics of similarity between color histograms. Assuming that the histograms are normalized ($\sum_j P(\omega_j) = 1$), the minimization of the L^1 distance is equivalent to the maximization of the HI [55]

$$\mathcal{D}_1(P||P_i) = \arg\max_i \sum_r \min[P(\omega_r), P_i(\omega_r)] \quad (14)$$

a similarity function that has become the *de facto* standard for color-based retrieval [1], [23], [32], [45], [48], [51], [53]–[55]. Since histograms have exponential complexity in the dimension of the feature space, and the ability to model the image dependencies that characterize texture usually requires high-dimensional feature vectors, an alternative set of similarity functions has evolved in the texture retrieval literature. In this literature, a popular representation is to summarize the pdf of the ith class by a few of its moments, typically the mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$, and rely on a quadratic distance

$$\mathcal{M}(P||P_i) = ||\boldsymbol{\mu} - \boldsymbol{\mu}_i||^2_{\boldsymbol{B}} \quad (15)$$

to compare it to those of the query, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The matrix $\boldsymbol{B}$ is usually the identity, or a function of the covariances $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_i$ [1], [4], [6], [23], [32], [34], [38], [42], [44], [45], [48], [50], [52]. It is clear from the comparison of (15) with (7) that the minimization of these distances is equivalent to the maximization of an approximation to the EL (where the terms that only depend on the covariances are dropped), under the assumption that all classes are Gaussian. This is a reasonable assumption when the images are homogeneous, a property that holds for most texture databases (whose images usually consist of uniform texture patches) but not for generic imagery.

More recently, there has been interest in the joint modeling of color and texture. One possibility to achieve this goal is by resorting to extensions of the histogram, that preserve information about the spatial relationships between colors. The most popular among such models is the autocorrelogram [20], a variant of the histogram, that includes the distance between pixels as an extra parameter. Besides the relative frequencies of the different colors, the autocorrelogram also stores the frequencies with which the colors occur simultaneously in pairs of pixels that are less than d image locations apart (where d is a free parameter). Another possibility is to rely on probabilistic models that provide estimates of the joint density of local image neighborhoods, e.g., a VQ. This has indeed been proposed by Rubner and colleagues [47] that also proposed the earth mover's distance (EMD) as a similarity metric for VQs. The EMD between two quantizers

$$P(\boldsymbol{x}) = \sum_k \delta(\boldsymbol{x} - \boldsymbol{\mu}_k)P(\omega_k)$$

and

$$Q(\boldsymbol{x}) = \sum_k \delta(\boldsymbol{x} - \boldsymbol{\eta}_k)Q(\omega_k)$$

is defined in terms of a flow matrix $\boldsymbol{F} = (f_{ij})$ that minimizes

$$W(P,Q,F) = \sum_{ij} f_{ij} d(\boldsymbol{\mu}_i, \boldsymbol{\eta}_j) \quad (16)$$

where $d(\cdot,\cdot)$ is a distance measure, under the constraints that the f_{ij} are nonnegative and

$$\sum_j f_{ij} = P(\omega_i)$$
$$\sum_i f_{ij} = Q(\omega_j)$$
$$\sum_{ij} f_{ij} = \min\left(\sum_i P(\omega_i), \sum_j Q(\omega_j)\right).$$

Given the optimal flow $\boldsymbol{F}^*$, the EMD is defined as

$$\mathcal{W}(P,Q) = \frac{\sum_{ij} f^*_{ij} d(\boldsymbol{\mu}_i, \boldsymbol{\eta}_j)}{\sum_{ij} f^*_{ij}}. \quad (17)$$

As pointed out in [28], it is not difficult to show that, when the VQs are normalized ($\sum_i P(\omega_i) = \sum_j Q(\omega_j) = 1$) the EMD is identical to the Wasserstein distance [65] between probability densities. Given two random variables A and B, with probability densities P and Q, respectively, the Wasserstein distance is the infimum of the expected distance between A and B, where the infimum is taken over the expectations with respect to all joint densities $F(A,B)$ that have marginal density P for A and Q for B

$$\mathcal{W}(P,Q) = \inf_F \{E_F[d(A,B)] \,|\, (A,B) \sim F, A \sim P, B \sim Q\}. \quad (18)$$

The Wasserstein distance has a long history in information theory and statistics, where it is also known as the rho-bar distance, the d-bar distance, the Ornstein distance, or the Mallows distance, see e.g., [15], [16], [27], [33], [39]. Vallender has shown [59] that when $d(\cdot,\cdot)$ is the L^1 metric

$$\mathcal{W}(P,Q) = \int |D_P(\boldsymbol{x}) - D_Q(\boldsymbol{x})| d\boldsymbol{x} \tag{19}$$

where D_P (D_Q) is the pdf associated with the pdf $P(\boldsymbol{x})$ $(Q(\boldsymbol{x}))$. Hence, the EMD is not fundamentally different HI: while the latter minimizes the L^1 distance between the probability densities, the former minimizes the L^1 distance between the associated distribution functions.

Given that these similarity functions either 1) have no explicit connection to the minimization of the probability of retrieval error, or 2) assume probability distributions, e.g., the Gaussian, that are unrealistic for image retrieval, it is questionable that they will lead to image retrieval systems that are optimal in a minimum-probability-of-error sense. This goal is, in principle, attainable by combining the MAP similarity function with less restrictive probabilistic models, such as the GM, that can approximate any pdf arbitrarily well [29]. However, the absence of closed-form solutions to the KL divergence between mixtures makes this solution too complex from a computational point of view, and applications of the KL divergence to image retrieval have been limited to histogram-based representations [5], [22], [45] (that have limited capacity as joint models of color and texture), or representations that assume independence between features [9] (a questionable assumption that has been shown not to work well in experimental studies [62]). In the following sections, we address the problem of computing the KL divergence between multivariate mixture models.

III. KL Divergence Between Vector Quantizers

We have seen in the previous section that the database density $P_i(\boldsymbol{x})$ closest to that of the query $P(\boldsymbol{x})$, in the ML sense, is that which maximizes the EL of $P_i(\boldsymbol{x})$ under $P(\boldsymbol{x})$

$$i^* = \arg\max_i \mathrm{EL}\,(P\|P_i) = \arg\max_i \int P(\boldsymbol{x}) \log P_i(\boldsymbol{x}) d\boldsymbol{x}. \tag{20}$$

For this reason, we will, from now on, simply refer to EL $(P\|P_i)$ as the retrieval similarity function. Furthermore, unless otherwise noted, all densities are assumed to be mixtures of the form (4). Given no constraint on the feature class-conditional densities $P(\boldsymbol{x}|\omega_j)$ and $P_i(\boldsymbol{x}|\omega_j)$, it is only possible to derive a generic expression for the similarity function.

Lemma 1: For a retrieval problem with query and database densities $P(\boldsymbol{x})$ and $P_i(\boldsymbol{x})$

$$\mathrm{EL}\,(P\|P_i) = \sum_{j,k} P(\omega_j) \int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x} \left[\log P_i(\omega_k) + \int_{\chi_k} P(\boldsymbol{x}|\omega_j, \chi_k(\boldsymbol{x})=1) \log \frac{P_i(\boldsymbol{x}|\omega_k)}{P_i(\omega_k|\boldsymbol{x})} d\boldsymbol{x} \right] \tag{21}$$

where

$$\chi_k(\boldsymbol{x}) = \begin{cases} 1, & \text{if } P_i(\omega_k|\boldsymbol{x}) \geq P_i(\omega_l|\boldsymbol{x}), \quad \forall l \neq k \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

$\chi_k = \{\boldsymbol{x} : \chi_k(\boldsymbol{x}) = 1\}$ defines the partition of the feature space, and

$$P(\boldsymbol{x}|\omega_j, \chi_k(\boldsymbol{x})=1) = \begin{cases} \frac{P(\boldsymbol{x}|\omega_j)}{\int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x}}, & \text{if } \boldsymbol{x} \in \chi_k \text{ and} \\ & \int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Proof: See Appendix II.

Equation (21) reveals two fundamental components of similarity. The first

$$\sum_{j,k} P(\omega_j) \log P_i(\omega_k) \int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x}$$

is a function of the feature class probabilities, the second

$$\sum_{j,k} P(\omega_j) \int_{\chi_k} P(\boldsymbol{x}|\omega_j) \log \frac{P_i(\boldsymbol{x}|\omega_k)}{P_i(\omega_k|\boldsymbol{x})} d\boldsymbol{x}$$

is a function of the class-conditional densities. The overall similarity is strongly dependent on the partition $\{\chi_1, \ldots, \chi_{C_i}\}$ of the feature space determined by the database density $P_i(\boldsymbol{x})$, the term

$$\int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x}$$

weighting the contribution of each cell according to the fraction of the query probability that it contains. In particular, if $\mathcal{S}(\omega_j)$ is the support set of $P(\boldsymbol{x}|\omega_j)$, then

$$\begin{aligned} &\int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x} = 0, && \text{if } \mathcal{S}(\omega_j) \cap \chi_k = \emptyset \\ &\int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x} = 1, && \text{if } \mathcal{S}(\omega_j) \subset \chi_k \\ &\int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x} \in (0,1), && \text{otherwise} \end{aligned}$$

and $\int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x}$ can be seen as a measure of overlap between $P(\boldsymbol{x}|\omega_j)$ and the cell χ_k determined by $P_i(\boldsymbol{x}|\omega_k)$.

A. Histograms

When all image classes share the same feature class-conditional densities and the partition of the feature space is fixed, determining the similarity in closed form is straightforward. This is the case of the histogram.

Lemma 2: If all mixture densities define the same hard partition

$$\chi_k(\boldsymbol{x}) = \begin{cases} 1, & \text{if } P(\omega_l|\boldsymbol{x}) = P_i(\omega_l|\boldsymbol{x}) = \delta_{k,l}, \quad \forall i \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where $\delta_{k,l}$ is the Kronecker delta function

$$\delta_{k,l} = \begin{cases} 1, & \text{if } k = l \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

then

$$\mathrm{EL}\,(P\|P_i) = \sum_j P(\omega_j) \log P_i(\omega_j) + \sum_j P(\omega_j) \int_{\chi_j} P(\boldsymbol{x}|\omega_j) \log P_i(\boldsymbol{x}|\omega_j) d\boldsymbol{x}.$$

Proof: See Appendix III.

Because, when all image classes share the same feature-class conditional densities, i.e., $P_l(\boldsymbol{x}|w_j) = P_m(\boldsymbol{x}|w_j) \, \forall (l,m)$, the second term of (25) does not depend on i, this lemma implies that

$$\arg\max_i \mathrm{EL}\,(P\|P_i) = \arg\max_i \sum_j P(\omega_j) \log P_i(\omega_j) = \arg\min_i \sum_j P(\omega_j) \log \frac{P(\omega_j)}{P_i(\omega_j)}$$

and we obtain the expression for the KL divergence between histograms (12).

B. Gauss Mixtures (GMs)

Lifting the restriction to a common hard partition makes the computation of the KL divergence significantly more challenging. We next concentrate on this case, starting with a preliminary result.

Lemma 3: For any probability density $P(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{\alpha} \in \mathbb{R}^n$, symmetric positive definite matrix $\boldsymbol{B} \in \mathbb{R}^{n \times n}$, and set χ, if

$$\int_\chi P(\boldsymbol{x}) d\boldsymbol{x} = 1$$

then

$$\int_\chi P(\boldsymbol{x}) \|\boldsymbol{x} - \boldsymbol{\alpha}\|^2_{\boldsymbol{B}} d\boldsymbol{x} = \mathrm{trace}[\boldsymbol{B}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}] + \|\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} - \boldsymbol{\alpha}\|^2_B,$$

where

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \int_\chi P(\boldsymbol{x}) \boldsymbol{x} \, d\boldsymbol{x}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \int_\chi P(\boldsymbol{x}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T d\boldsymbol{x}.$$

Proof: See Appendix IV.

This lemma allows us to specialize (21) to GMs.

Lemma 4: For a retrieval problem with query density $P(\boldsymbol{x})$ given by (4) and GMs for the database densities $P_i(\boldsymbol{x})$

$$P_i(\boldsymbol{x}) = \sum_{k=1}^{C_i} \mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}) P_i(\omega_k)$$

where $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is as defined in (5)

$$\begin{aligned} \mathrm{EL}(P\|P_i) = \sum_{j,k} P(\omega_j) \int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x} \Bigg\{ & \log P_i(\omega_k) \\ & + \left[\log \mathcal{G}(\hat{\boldsymbol{\mu}}_{q,j,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}) - \frac{1}{2} \mathrm{trace}[\boldsymbol{\Sigma}_{i,k}^{-1} \hat{\boldsymbol{\Sigma}}_{q,j,k}] \right] \\ & - \int_{\chi_k} P(\boldsymbol{x}|\omega_j, \chi_k(\boldsymbol{x}) = 1) \log P_i(\omega_k|\boldsymbol{x}) d\boldsymbol{x} \Bigg\} \end{aligned} \tag{25}$$

where

$$\hat{\boldsymbol{\mu}}_{q,j,k} = \int_{\chi_k} P(\boldsymbol{x}|\omega_j, \chi_k(\boldsymbol{x}) = 1) \boldsymbol{x} \, d\boldsymbol{x} \tag{26}$$

$$\hat{\boldsymbol{\Sigma}}_{q,j,k} = \int_{\chi_k} P(\boldsymbol{x}|\omega_j, \chi_k(\boldsymbol{x}) = 1)(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{q,j,k})(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{q,j,k})^T d\boldsymbol{x} \tag{27}$$

and χ_k and $P(\boldsymbol{x}|\omega_j, \chi_k(\boldsymbol{x}) = 1)$ are as defined in Lemma 1.

Proof: See Appendix V.

Equation (25) reveals that, for GMs, there are three components to the similarity function. Consider, without loss of generality, the query feature subclass w_j and the database feature class w_k. The first term in the equation is simply a measure of the similarity between the class probabilities $P(\omega_j)$ and $P_i(\omega_k)$ weighted by the measure of overlap $\int_{\chi_k} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x}$. This term is a generalization of the one appearing in (12) that accounts for the lack of alignment between the partitions defined by the query and database densities.

The term in square brackets is, up to constants that do not depend on i, the KL divergence between the Gaussian $P_i(\boldsymbol{x}|\omega_k)$ and a Gaussian with parameters $\hat{\boldsymbol{\mu}}_{q,j,k}$ and $\hat{\boldsymbol{\Sigma}}_{q,j,k}$ [25]. From (26) and (27), these are simply the mean and covariance of $\boldsymbol{x}$ under $P(\boldsymbol{x}|\omega_j)$ given that $\boldsymbol{x} \in \chi_k$. Hence, the second term is simply a measure of the similarity between the feature class conditional densities inside the cell defined by $P_i(\boldsymbol{x}|\omega_k)$. Once again, this measure is weighted by the amount of overlap between the two densities.

Finally, the third term weighs the different cells χ_k according to the ambiguity of their ownership. Recall that, $\forall \boldsymbol{x} \in \chi_k$, $P_i(\omega_k|\boldsymbol{x}) > P_i(\omega_l|\boldsymbol{x})$, $\forall l \neq k$. If $P_i(\omega_k|\boldsymbol{x}) = 1$, the cell is uniquely assigned to ω_k and this term will be zero. If, on the other hand, $P_i(\omega_k|\boldsymbol{x}) < 1$, then the cell will also be assigned to other classes and the overall similarity will increase.

While providing insight on the different components of similarity, (25) is not very useful from a computational standpoint since the integrals that it involves do not have a closed-form solution. There is, however, one particular case where such a solution exits: the case where all mixture models are VQs.

C. Vector Quantizers (VQs)

Using Theorem 2, the VQ case can be analyzed by assuming Gaussian feature class-conditional densities and investigating what happens when all covariance matrices tend to zero. This leads to the following result.

Lemma 5: For a retrieval problem with GMs for the query and database densities

$$P_\epsilon(\boldsymbol{x}) = \sum_{j=1}^{C} \mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_{q,j}, \epsilon\boldsymbol{\Sigma}_{q,j}) P(\omega_j)$$

$$P_{i,\epsilon}(\boldsymbol{x}) = \sum_{k=1}^{C_i} \mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_{i,k}, \epsilon\boldsymbol{\Sigma}_{i,k}) P_i(\omega_k)$$

where $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is as defined in (5), when $\epsilon \to 0$

$$\begin{aligned} \mathrm{EL}\,(P_\epsilon\|P_{i,\epsilon}) = \sum_j P(\omega_j) \Bigg\{ & \log P_i(\omega_{\alpha(j)}) \\ & + \lim_{\epsilon \to 0} \Bigg[\log \mathcal{G}(\hat{\boldsymbol{\mu}}_{q,j,\alpha(j)}, \boldsymbol{\mu}_{i,\alpha(j)}, \epsilon\boldsymbol{\Sigma}_{i,\alpha(j)}) \\ & - \frac{1}{2\epsilon} \mathrm{trace}[\boldsymbol{\Sigma}_{i,\alpha(j)}^{-1} \hat{\boldsymbol{\Sigma}}_{q,j,\alpha(j)}] \Bigg] \Bigg\} \end{aligned}$$

where χ_k is as defined in Lemma 1, $\hat{\boldsymbol{\mu}}_{q,j,\alpha(j)}$ and $\hat{\boldsymbol{\Sigma}}_{q,j,\alpha(j)}$ as defined in Lemma 4, and

$$\alpha(j) = k \iff \|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,k}\|^2_{\boldsymbol{\Sigma}_{i,k}} < \|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,l}\|^2_{\boldsymbol{\Sigma}_{i,l}} \, \forall l \neq k.$$

Proof: See Appendix VI.

We are now ready to derive a closed-form expression for the similarity between VQ-based density estimates.

Theorem 3: For a retrieval problem with VQ-based estimates for the query and database densities

$$P(\boldsymbol{x}) = \sum_{j=1}^{C} \delta(\boldsymbol{x} - \boldsymbol{\mu}_{q,j}) P(\omega_j)$$

$$P_i(\boldsymbol{x}) = \sum_{k=1}^{C_i} \delta(\boldsymbol{x} - \boldsymbol{\mu}_{i,k}) P_i(\omega_k)$$

the KL similarity criteria reduce to

$$\begin{aligned}\arg\max_i \mathrm{EL}\,(P\|P_i) &= \arg\min_i \lim_{\lambda\to\infty} \left\{ \sum_j P(\omega_j) \log \frac{P(\omega_j)}{P_i(\omega_{\alpha(j)})} \right. \\ &\qquad \left. + \lambda \sum_j P(\omega_j) \|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,\alpha(j)}\|^2 \right\} \quad (28)\end{aligned}$$

where

$$\alpha(j) = k \iff \|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,k}\|^2 < \|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,l}\|^2, \forall l \neq k.$$

Proof: See Appendix VII.

The theorem states that, for VQ-based density estimates, the minimization of the KL divergence is a constrained optimization problem [3]. Given a query VQ and a database VQ, one starts by vector quantizing the codewords of the former according to the latter, i.e., each codeword of the query VQ is assigned to the cell of the database VQ whose centroid is closest to it. The best database VQ is the one that minimizes a sum of two resulting terms: a term that accounts for the average distortion of the quantization ($\sum_j P(\omega_j)\|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,\alpha(j)}\|^2$) and the KL divergence between the feature-class probability distributions. λ is a Lagrange multiplier that weighs the contribution of the two terms. By taking the limit $\lambda \to \infty$, all the emphasis is placed on the average quantization distortion. This leads to two distinct situations of practical interest. The first is when the two quantizers share the same codewords. In this case, the quantization distortion is null and the cost function becomes that of (12), i.e., the KL divergence between label histograms. Since equal quantizers with equal codewords define equal partitions of the feature space, this situation is equivalent to that of histograms and the result is, therefore, not surprising.

The second is when the quantizers have different codewords (and consequently define different partitions). In this case, the quantization distortion becomes predominant and the retrieval criteria reduces to

$$\arg\max_i \mathrm{EL}\,(P\|P_i) = \arg\min_i \sum_j P(\omega_j)\|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,\alpha(j)}\|^2.$$

Note that, even in this case, the complexity of the retrieval operation is only $O(C^2 n)$, where n the dimension of the space. $C^2 n$ is typically orders of magnitude smaller than the cardinality of the query set Q leading to significant savings over the direct evaluation of the query likelihood. Compared to the complexity of histogram-based techniques $O(b^n)$, where b is the number of bins per axis, (28) trades off exponential growth in the number of dimensions by quadratic growth in the number of classes. Since C is usually small, this enables significantly more accurate estimates in high-dimensional spaces. Consider, for example, a space with $n = 16$: the complexity of (28) with 64 mixture components (more than enough to accommodate the typical number of clusters in densities of practical interest) is equal to the complexity of HI with only two bins per axis (i.e., each feature quantized in a binary fashion). It is natural to expect that the coarse histogram estimates will lead to worse retrieval performance.

IV. The Asymptotic Likelihood Approximation

Vector quantization has particular interest not only because it leads to a closed-form similarity expression, but also because it provides insights on how to approximate (25) in the case of generic GMs. In particular, Lemma 5 suggests the following approximation.

Definition 1: Given a retrieval problem with GMs for the query and database densities

$$P(\boldsymbol{x}) = \sum_{j=1}^{C} \mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_{q,j}, \boldsymbol{\Sigma}_{q,j}) P(\omega_j)$$

$$P_i(\boldsymbol{x}) = \sum_{k=1}^{C_i} \mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}) P_i(\omega_k)$$

the ALA is defined as

$$\begin{aligned}\mathrm{ALA}\,(P\|P_i) = \sum_j P(\omega_j) &\left\{ \log P_i(\omega_{\beta(j)}) \right. \\ &\left. + \left[\log \mathcal{G}(\boldsymbol{\mu}_{q,j}, \boldsymbol{\mu}_{i,\beta(j)}, \boldsymbol{\Sigma}_{i,\beta(j)}) - \frac{1}{2}\mathrm{trace}[\boldsymbol{\Sigma}_{i,\beta(j)}^{-1}\boldsymbol{\Sigma}_{q,j}] \right] \right\}\end{aligned}$$

where

$$\begin{aligned}\beta(j) = k \iff &\|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,k}\|^2_{\boldsymbol{\Sigma}_{i,k}} - \log P_i(\omega_k) \\ &< \|\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,l}\|^2_{\boldsymbol{\Sigma}_{i,l}} - \log P_i(\omega_l), \qquad \forall l \neq k. \quad (29)\end{aligned}$$

The following theorem provides the conditions under which the approximation is exact.

Theorem 4: For the retrieval problem of Definition 1

$$\mathrm{ALA}\,(P\|P_i) = \mathrm{EL}\,(P\|P_i) \quad (30)$$

when the following two conditions hold.

1) Each cell χ_k of the partition determined by $P_i(\boldsymbol{x})$ is assigned to one feature class with probability one, i.e.,

$$P_i(\omega_k|\boldsymbol{x}) = 1, \qquad \forall \boldsymbol{x} \in \chi_k. \quad (31)$$

2) The support set of each feature class-conditional density of the query mixture is entirely contained in a single cell χ_k of the partition determined by $P_i(\boldsymbol{x})$, i.e.,

$$\forall j, \qquad \exists k : \mathcal{S}(\omega_j) \subset \chi_k. \quad (32)$$

Proof: See Appendix VIII.

In a strict sense, the conditions of the theorem only hold for the VQ case since, when covariances are nonzero, the Gaussian class-conditional densities have unbounded support and none of the two conditions can be met. However, (31) will still hold approximately if the distance between each pair of mean vectors $\boldsymbol{\mu}_{i,k}$ is significantly larger than the spread of the associated Gaussians. A one-dimensional (1-D) illustration of this effect

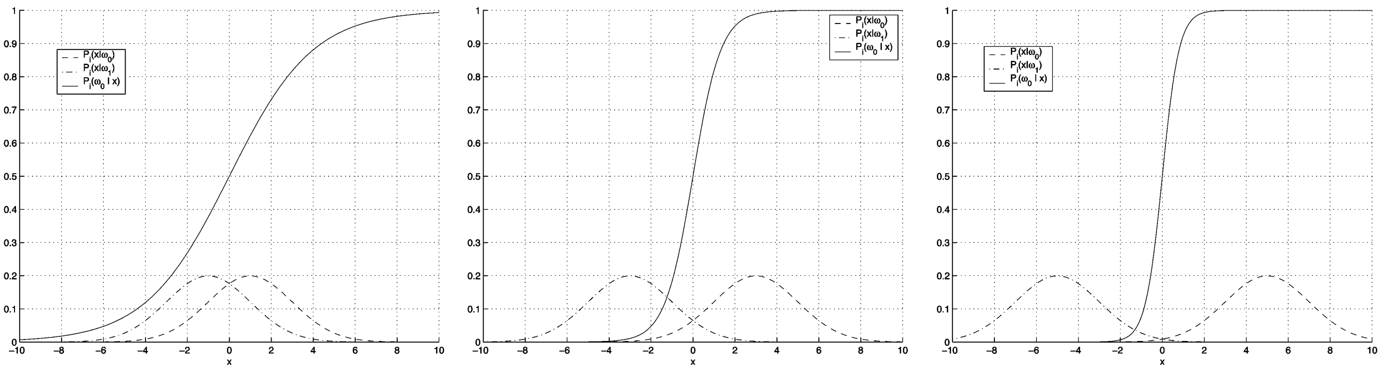


Fig. 1. Impact of the separation between two Gaussian densities (dashed) on the partition that they define (solid line).

is provided in Fig. 1, where we show two Gaussian class-conditional likelihood functions and the posterior probability function $P_i(\omega_0|\boldsymbol{x})$ for class 0. As the separation between the Gaussians increases, the posterior probability changes more abruptly and the partition becomes harder. If, in addition to this reduced overlap between the mixture components of the database density, the spread of the Gaussians in the query density $P(\boldsymbol{x})$ is much smaller than the size of the cells χ_k, then

$$\int_{\chi_\beta(j)} P(\boldsymbol{x}|\omega_j)d\boldsymbol{x} \approx 1$$

with high probability and condition 2) is also met approximately. In summary, the crucial assumption for the validity of the ALA is that the Gaussian feature class-conditional densities within each model have reduced overlap. The plausibility of this assumption grows with the dimension of the feature space, since high-dimensional spaces are more sparsely populated than low-dimensional ones. In the next section, we provide experimental evidence in support of this argument.

V. Experimental Evaluation

We performed two sets of experiments to evaluate the performance of probabilistic image retrieval with both the MAP and ALA similarity functions. The first set was designed to test the argument that the latter is a good approximation to the former in high-dimensional spaces. The second was designed to evaluate the retrieval performance of both similarity functions in a real image retrieval task, and compare it to those of some previously proposed approaches that are popular in the CBIR literature. All experiments were performed on the Corel image database, and the feature space was the space of coefficients of the 8 × 8 discrete cosine transform (DCT) commonly used in image compression [41].

A. The Accuracy of the ALA

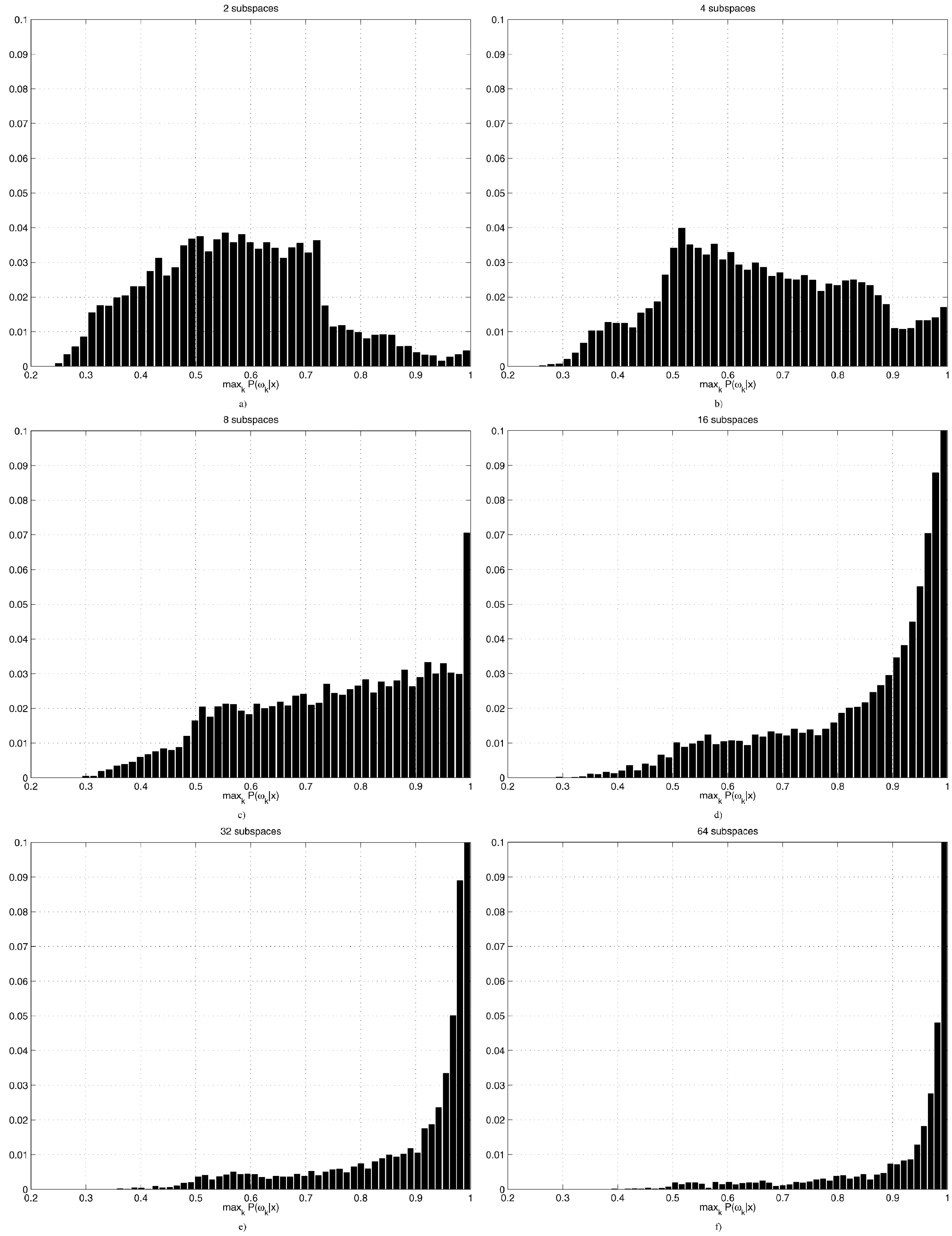
To evaluate how the accuracy of the ALA varies with the dimensionality of the feature space, we relied on the following Monte Carlo experiment:

- a test image was selected randomly from Corel and a training sample obtained by extracting DCT coefficients with a running window (moved over the image with increments of two pixels in a raster-scan fashion);
- the ML parameters of a GM with eight feature subclasses were computed from this sample, using the expectation–maximization algorithm [7];
- a sample with 10 000 points was drawn from this mixture model;
- for each sample point $\boldsymbol{x}_i, i = 1, \ldots, 10\,000$, the maximum posterior class-assignment probability $max_k P(\omega_k|\boldsymbol{x}_i)$ was computed;
- the maximum posterior probabilities were histogramed.

This procedure was repeated for different space dimensions, by projecting the mixture model of the 64-dimensional space into lower dimensional subspaces, and the whole experiment repeated with various images. Fig. 2 presents the histograms of the maximum posterior probability for 2, 4, 8, 16, 32, and 64 subspaces. It is clear that, as the dimension of the space increases, the probability that each sample is assigned to a single feature subclass increases. For example, when $n = 64$, the probability of this event is already close to 90%. This supports the argument that, for GMs in high-dimensional spaces, it is reasonable to assume that (31) holds.

B. Image Retrieval

In this subsection, we report the results of experiments on a dataset containing 1500 images from 15 classes of the Corel database, comparing the performance of probabilistic retrieval against HI and color autocorrelograms. In these experiments, we have used mixtures of eight Gaussians on a 48-dimensional feature space consisting of the 16 lower frequency DCT coefficients from each color channel (see [60], [64] for details), 512-bin color histograms, and 2048-bin color autocorrelograms (512-bin base histograms times four distances, see [20] for details). While, to be completely fair (exact same amount of computation), we should have used 3072 bins for the histogram techniques, preliminary experiments had shown no performance increase over the results obtained with the number of parameters above. In order to evaluate the goodness of the ALA, we evaluated the retrieval performance under both MALA (where we maximize the ALA) and the, much more expensive, exact ML similarity function. The plot on Fig. 3 presents the precision/recall curves for the four different retrieval solutions. It is clear that the performance of HI is not very good, and autocorrelo-



a)

b)

c)

d)

e)

f)

Fig. 2. Maximum class posterior probability histograms illustrate how the overlap between feature subclasses decreases in high dimensions. Space dimensions: a) 2, b) 4, c) 8, d) 16, e) 32, f) 64). Note that plots d)–f) are clipped at 0.1 in the vertical axis to allow the visualization of the histogram tails.

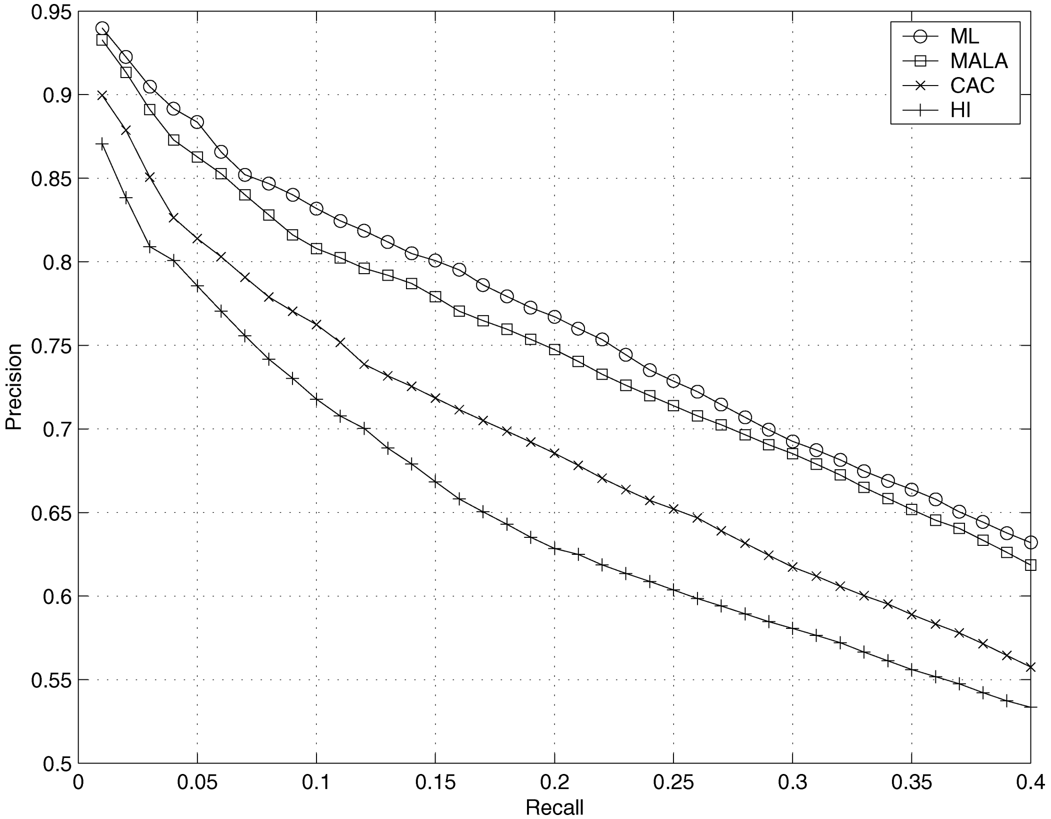


Fig. 3. Precision/recall on Corel for HI, color autocorrelogram (CAC), ML, and MALA.

grams only improve performance by about 5%. Both approaches are significantly less effective than either ML or MALA, that show no significant differences in precision/recall. This supports the argument that, 1) ALA is a good approximation to the true likelihood, and 2) MALA is the best overall solution when one takes computational complexity into account. We conclude by presenting some visual examples of the retrieval outcome. Fig. 4 presents typical results for queries with horses, cars, and oil paintings. These results illustrate some of the nice properties of the probabilistic retrieval formulation: robustness to changes in object position and orientation, robustness against the presence of distracting objects in the background, and perceptually intuitive errors (in the painting example, two pictures of the sphinx—pyramids class—are returned after all the paintings of human figures).

VI. Conclusion

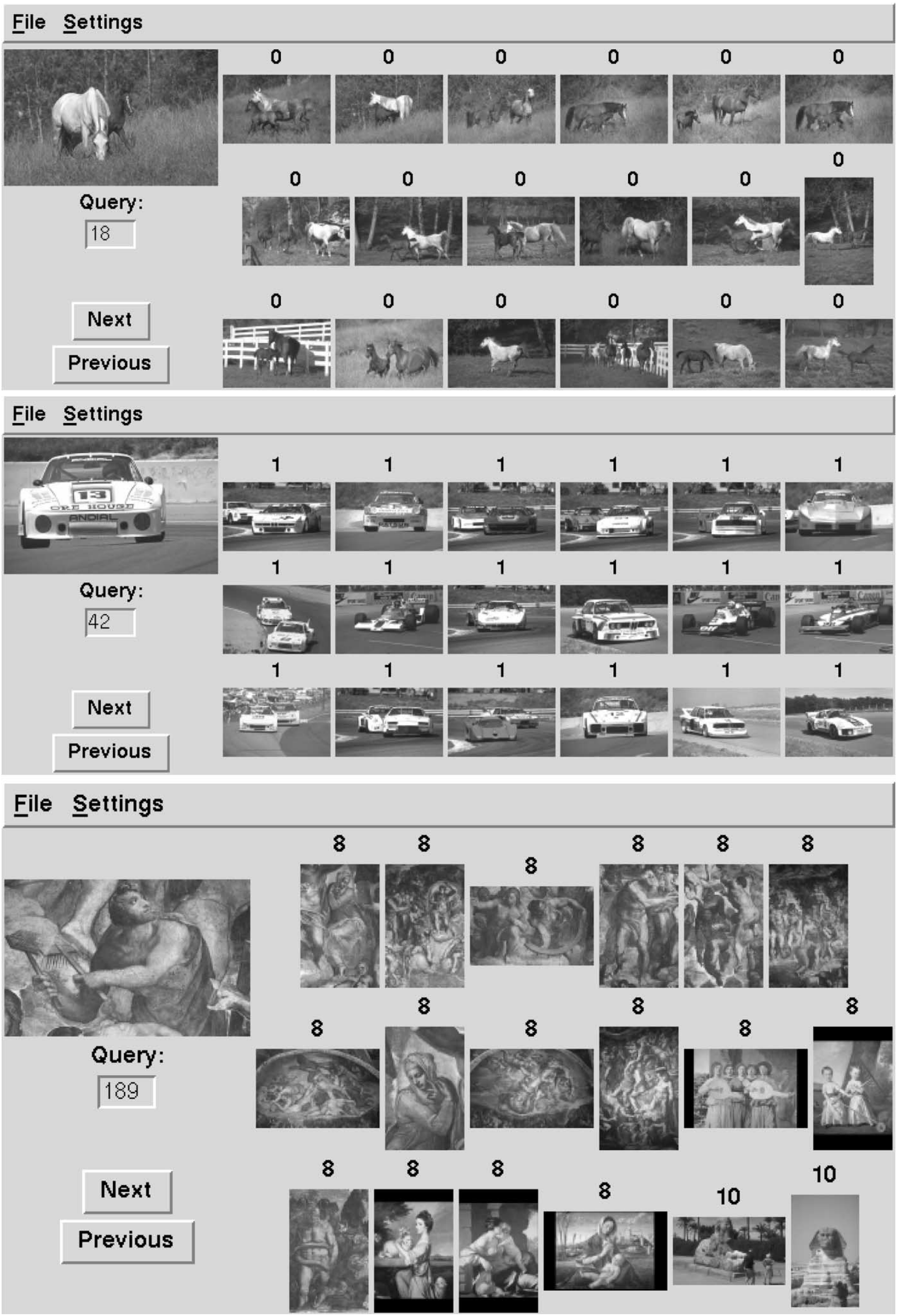
Probabilistic solutions have shown great promise for the CBIR problem, but have traditionally been difficult to deploy in practice due to the complexity of the MAP similarity function. In this work, we have shown that this similarity function can be computed efficiently when VQs are used as models for the pdfs of the image features. We have also argued that in the more general GM case, the MAP function can be well approximated by a closed-form expression of reduced complexity, which we denoted by ALA. The accuracy of this approximation increases with the dimension of the feature space in which similarity is computed. Experimental evaluation has shown that the combination of probabilistic retrieval (using either of these functions) with GMs can outperform retrieval techniques that are popular in the area of CBIR.

While the focus of this work has been in image retrieval, we believe that the conclusions will hold for any other databases containing signals characterized by high-dimensional features. This includes music, audio, video, and even signals that have traditionally not been considered in the database literature, such as those associated with DNA information. We also believe that the theoretical significance of the results here presented goes well beyond the information retrieval problem. Mixture models are widely used in statistical modeling and computing the KL divergence between them is a problem that appears frequently when they are employed. Currently, this implies the use of Monte Carlo procedures that are computationally expensive. The complexity can be overwhelming when this process has to be repeated within some other algorithm, e.g., expectation–maximization or inference algorithms for probabilistic networks [24]. We believe that the results presented in this work can dramatically improve the computational efficiency of such algorithms.

Appendix I
Proof of Theorem 2

Since a mixture model with C classes of which z have zero probability is the same as a model with $C-z$ classes of nonzero probability, we assume, without loss of generality, that all the classes have nonzero probability, i.e.,

$$P(\omega_i) > 0, \qquad \forall i.$$



Query: 18

Query: 42

Query: 189

Fig. 4. Queries for horses, cars, and paintings.

For Gaussian feature class-conditional densities, (8) then becomes

$$P(\omega_i|\boldsymbol{x}) = \frac{1}{1+\sum_{k\neq i}\sqrt{\frac{|\boldsymbol{\Sigma}_i|}{|\boldsymbol{\Sigma}_k|}}\,\frac{e^{||\boldsymbol{x}-\boldsymbol{\mu}_i||^2_{\boldsymbol{\Sigma}_i}-\log P(\omega_i)}}{e^{||\boldsymbol{x}-\boldsymbol{\mu}_k||^2_{\boldsymbol{\Sigma}_k}-\log P(\omega_k)}}}$$

and for $\boldsymbol{\Sigma}_i = \epsilon\boldsymbol{I}, \forall i$

$$P_\epsilon(\omega_i|\boldsymbol{x}) = \frac{1}{1+\sum_{k\neq i}\frac{P(\omega_k)}{P(\omega_i)}e^{\frac{1}{\epsilon}(||\boldsymbol{x}-\boldsymbol{\mu}_i||^2-||\boldsymbol{x}-\boldsymbol{\mu}_k||^2)}}.$$

Hence,

$$\lim_{\epsilon\longrightarrow 0} P_\epsilon(\omega_i|\boldsymbol{x}) = \begin{cases} a, & \text{if } ||\boldsymbol{x}-\boldsymbol{\mu}_i||^2 \leq ||\boldsymbol{x}-\boldsymbol{\mu}_k||^2,\ \forall k\neq i \\ 0, & \text{otherwise.,} \end{cases} \tag{33}$$

where

$$a = \frac{P(\omega_i)}{P(\omega_i)+\sum_{\{k:||\boldsymbol{x}-\boldsymbol{\mu}_k||=||\boldsymbol{x}-\boldsymbol{\mu}_i||\}}P(\omega_k)}.$$

Since the set $\{\boldsymbol{x} : ||\boldsymbol{x}-\boldsymbol{\mu}_k|| = ||\boldsymbol{x}-\boldsymbol{\mu}_i||\}$ has measure zero, (33) is equivalent to (9) almost everywhere. Furthermore, because some arbitrary tie-breaking rule is always necessary to vector-quantize the points that lie on the boundaries between different cells, the same rule can be applied to (33) and the two equations are equivalent. Equation (10) is a direct consequence of the fact that the Gaussian density converges to the delta function as its covariance tends to zero [40]. □

Appendix II
Proof of Lemma 1

From (4) and (2)

$$\begin{aligned}\text{EL}\,(P||P_i) &= \sum_j P(\omega_j)\int P(\boldsymbol{x}|\omega_j)\log\sum_l P_i(\boldsymbol{x}|\omega_l)P_i(\omega_l)d\boldsymbol{x} \\ &= \sum_j P(\omega_j)\sum_k\int_{\chi_k} P(\boldsymbol{x}|\omega_j)\log\sum_l P_i(\boldsymbol{x}|\omega_l)P_i(\omega_l)d\boldsymbol{x}.\end{aligned}$$

Using Bayes rule

$$P_i(\omega_k|\boldsymbol{x}) = \frac{P_i(\boldsymbol{x}|\omega_k)P_i(\omega_k)}{\sum_l P_i(\boldsymbol{x}|\omega_l)P_i(\omega_l)} \tag{34}$$

we have $\forall k$ such that $P_i(\omega_k|\boldsymbol{x}) \neq 0$

$$\sum_l P_i(\boldsymbol{x}|\omega_l)P_i(\omega_l)d\boldsymbol{x} = \frac{P_i(\boldsymbol{x}|\omega_k)P_i(\omega_k)}{P_i(\omega_k|\boldsymbol{x})}.$$

Since $\sum_k P_i(\omega_k|\boldsymbol{x}) = 1$, from the definition of χ_k we obtain $P_i(\omega_k|\boldsymbol{x}) > 0$, $\forall\boldsymbol{x}\in\chi_k$, and

$$\begin{aligned}El(P||P_i) &= \sum_j P(\omega_j)\sum_k\int_{\chi_k} P(\boldsymbol{x}|\omega_j)\log\frac{P_i(\boldsymbol{x}|\omega_k)P_i(\omega_k)}{P_i(\omega_k|\boldsymbol{x})}d\boldsymbol{x} \\ &= \sum_j P(\omega_j)\sum_k\Bigg[\log P_i(\omega_k)\int_{\chi_k} P(\boldsymbol{x}|\omega_j)d\boldsymbol{x} \\ &\quad + \int_{\chi_k} P(\boldsymbol{x}|\omega_j)\log\frac{P_i(\boldsymbol{x}|\omega_k)}{P_i(\omega_k|\boldsymbol{x})}d\boldsymbol{x}\Bigg] \\ &= \sum_j P(\omega_j)\sum_k\int_{\chi_k} P(\boldsymbol{x}|\omega_j)d\boldsymbol{x}\,\Bigg[\log P_i(\omega_k) \\ &\quad + \int_{\chi_k} P(\boldsymbol{x}|\omega_j,\chi_k(\boldsymbol{x})=1)\log\frac{P_i(\boldsymbol{x}|\omega_k)}{P_i(\omega_k|\boldsymbol{x})}d\boldsymbol{x}\Bigg].\ \square\end{aligned}$$

Appendix III
Proof of Lemma 2

Using the same argument as in the proof of Theorem 2, we assume, without loss of generality, that all the classes in all mixture models have nonzero probability, i.e.,

$$P(\omega_l) > 0 \ \text{ and } P_i(\omega_l) > 0, \qquad \forall l, i.$$

From (34), $P_i(\omega_k|\boldsymbol{x}) = 1$ if and only if

$$\sum_{l\neq k} P_i(\boldsymbol{x}|\omega_l)P_i(\omega_l) = 0$$

and since all the terms in the summation are nonnegative, this implies

$$P_i(\boldsymbol{x}|\omega_l) = 0 \qquad \forall l\neq k$$

i.e., for a hard partition such as (23), the support sets of $P(\boldsymbol{x}|\omega_k)$ and $P_i(\boldsymbol{x}|\omega_k)$ are contained in χ_k. Hence,

$$\int_{\chi_k} P(\boldsymbol{x}|\omega_j)d\boldsymbol{x} = \delta_{k,j}$$

and, since $P_i(\omega_k|\boldsymbol{x}) = 1, \forall\boldsymbol{x}\in\chi_k$, (25) follows from Lemma 1. □

Appendix IV
Proof of Lemma 3

$$\begin{aligned}\int_\chi P(\boldsymbol{x})||\boldsymbol{x}-\boldsymbol{\alpha}||^2_{\boldsymbol{B}}d\boldsymbol{x} &= \int_\chi P(\boldsymbol{x})||\boldsymbol{x}-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}+\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}-\boldsymbol{\alpha}||^2_{\boldsymbol{B}}d\boldsymbol{x} \\ &= \int_\chi P(\boldsymbol{x})||\boldsymbol{x}-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}||^2_{\boldsymbol{B}}d\boldsymbol{x} + ||\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}-\boldsymbol{\alpha}||^2_{\boldsymbol{B}} \\ &\quad + 2\int_\chi P(\boldsymbol{x})(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T\boldsymbol{B}^{-1}(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}-\boldsymbol{\alpha})d\boldsymbol{x} \\ &= \text{trace}\left[\boldsymbol{B}^{-1}\int_\chi P(\boldsymbol{x})(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T d\boldsymbol{x}\right] \\ &\quad + ||\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}-\boldsymbol{\alpha}||^2_{\boldsymbol{B}} + 2(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T\boldsymbol{B}^{-1}(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}-\boldsymbol{\alpha}) \\ &= \text{trace}[\boldsymbol{B}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}] + ||\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}-\boldsymbol{\alpha}||^2_{\boldsymbol{B}}.\qquad\square\end{aligned}$$

Appendix V
Proof of Lemma 4

Since $P_i(\boldsymbol{x}|\omega_k) = \mathcal{G}(\boldsymbol{x},\boldsymbol{\mu}_{i,k},\boldsymbol{\Sigma}_{i,k})$ and $\int_{\chi_k} P(\boldsymbol{x}|\omega_j,\chi_k(\boldsymbol{x}) = 1)d\boldsymbol{x} = 1$, simple application of the previous lemma results in

$$\begin{aligned}&\int_{\chi_k} P(\boldsymbol{x}|\omega_j,\chi_k(\boldsymbol{x})=1)\log P_i(\boldsymbol{x}|\omega_k)d\boldsymbol{x} = \\ &\quad = \log\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}_{i,k}|}}\int_{\chi_k} P(\boldsymbol{x}|\omega_j,\chi_k(\boldsymbol{x})=1)d\boldsymbol{x} \\ &\quad\quad - \frac{1}{2}\int_{\chi_k} P(\boldsymbol{x}|\omega_j,\chi_k(\boldsymbol{x})=1)||\boldsymbol{x}-\boldsymbol{\mu}_{i,k}||^2_{\boldsymbol{\Sigma}_{i,k}}d\boldsymbol{x} \\ &\quad = \log\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}_{i,k}|}} - \frac{1}{2}\text{trace}[\boldsymbol{\Sigma}^{-1}_{i,k}\hat{\boldsymbol{\Sigma}}_{q,j,k}] \\ &\quad\quad - \frac{1}{2}||\hat{\boldsymbol{\mu}}_{q,j,k}-\boldsymbol{\mu}_{i,k}||^2_{\boldsymbol{\Sigma}_{i,k}} \\ &\quad = \log\mathcal{G}(\hat{\boldsymbol{\mu}}_{q,j,k},\boldsymbol{\mu}_{i,k},\boldsymbol{\Sigma}_{i,k}) - \frac{1}{2}\text{trace}[\boldsymbol{\Sigma}^{-1}_{i,k}\hat{\boldsymbol{\Sigma}}_{q,j,k}].\end{aligned}$$

The lemma follows by simple algebraic manipulation of (21). □

APPENDIX VI
PROOF OF LEMMA 5

When $\epsilon \rightarrow 0$

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_{q,j}, \epsilon\boldsymbol{\Sigma}_{q,j}) \rightarrow \delta(\boldsymbol{x} - \boldsymbol{\mu}_{q,j})$$

and since, from the definition of the delta function

$$\int f(\boldsymbol{x})\delta(\boldsymbol{x} - \boldsymbol{\mu})d\boldsymbol{x} = f(\boldsymbol{\mu})$$

it follows that

$$\int_{\chi_k} P_\epsilon(\boldsymbol{x}|\omega_j)d\boldsymbol{x} \rightarrow \chi_k(\boldsymbol{\mu}_{q,j}).$$

On the other hand, from Theorem 2 and the definition of χ_k, if $\epsilon \rightarrow 0$ then

$$P_{i,\epsilon}(\omega_k|\boldsymbol{x}) \rightarrow 1, \qquad \forall \boldsymbol{x} \in \chi_k,$$

and $\chi_k(\boldsymbol{\mu}_{q,j}) = 1$ if and only if

$$||\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,k}||^2_{\boldsymbol{\Sigma}_{i,k}} < ||\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,l}||^2_{\boldsymbol{\Sigma}_{i,l}}, \qquad \forall l \neq k.$$

The lemma follows from the application of these results to (25). □

APPENDIX VII
PROOF OF THEOREM 3

From (26) and (27), when $\epsilon \rightarrow 0$

$$\hat{\boldsymbol{\mu}}_{q,j,\alpha(j)} \rightarrow \boldsymbol{\mu}_{q,j}$$
$$\hat{\boldsymbol{\Sigma}}_{q,j,\alpha(j)} \rightarrow \epsilon\boldsymbol{\Sigma}_{q,j}.$$

Using Lemma 5

$$\begin{aligned}\arg\max_i \mathrm{EL}\,(P||P_i) &= \arg\max_i \sum_j P(\omega_j)\Bigg\{\log P_i(\omega_{\alpha(j)}) - \frac{1}{2}\mathrm{trace}[\boldsymbol{\Sigma}^{-1}_{i,\alpha(j)}\boldsymbol{\Sigma}_{q,j}] \\ &\quad + \lim_{\epsilon\rightarrow 0}\log\mathcal{G}(\boldsymbol{\mu}_{q,j}, \boldsymbol{\mu}_{i,\alpha(j)}, \epsilon\boldsymbol{\Sigma}_{i,\alpha(j)})\Bigg\}.\end{aligned}$$

Since, for a VQ, $\boldsymbol{\Sigma}_{i,k} = \boldsymbol{\Sigma}_{q,j} = \boldsymbol{I}, \forall k, j$, the third term on the right-hand side of the preceding equation does not depend on i, and setting $\lambda = 1/2\epsilon$ leads to

$$\begin{aligned}\arg\max_i \mathrm{EL}\,(P||P_i) &= \arg\max_i \left\{\sum_j P(\omega_j)\Bigg[\log P_i(\omega_{\alpha(j)}) \right. \\ &\quad \left. - \lim_{\lambda\rightarrow\infty}\lambda||\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,\alpha(j)}||^2\Bigg]\right\} \\ &= \arg\min_i \lim_{\lambda\rightarrow\infty}\left\{\sum_j P(\omega_j)\log\frac{P(\omega_j)}{P_i(\omega_{\alpha(j)})}\right. \\ &\quad \left. + \lambda\sum_j P(\omega_j)||\boldsymbol{\mu}_{q,j} - \boldsymbol{\mu}_{i,\alpha(j)}||^2\right\}. \quad \square\end{aligned}$$

APPENDIX VIII
PROOF OF THEOREM 4

When condition 1) holds, the third term of (25) vanishes and

$$\begin{aligned}\mathrm{EL}(P||P_i) &= \sum_{j,k} P(\omega_j)\int_{\chi_k} P(\boldsymbol{x}|\omega_j)d\boldsymbol{x}\Bigg\{\log P_i(\omega_k) \\ &\quad + \left[\log\mathcal{G}(\hat{\boldsymbol{\mu}}_{q,j,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}) - \frac{1}{2}\mathrm{trace}[\boldsymbol{\Sigma}^{-1}_{i,k}\hat{\boldsymbol{\Sigma}}_{q,j,k}]\right]\Bigg\}.\end{aligned} \tag{35}$$

Since, from condition 2) $\mu_{q,j} \in \chi_k$, it follows from (22) and (8) that

$$P_i(\boldsymbol{\mu}_{q,j}|\omega_k)P_i(\omega_k) > P_i(\boldsymbol{\mu}_{q,j}|\omega_l)P_i(\omega_l), \qquad \forall l \neq k.$$

Taking logarithms on both sides and using the fact that

$$P_i(\boldsymbol{x}|\omega_m) = \mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_{i,m}, \boldsymbol{\Sigma}_{i,m})$$

leads, after some algebraic manipulation, to (29). It follows that $\mathcal{S}(\omega_j) \subset \chi_{\beta(j)}$ and

$$\int_{\chi_k} P(\boldsymbol{x}|\omega_j)d\boldsymbol{x} = \delta_{k,\beta(j)}. \tag{36}$$

Using the definition of $P(\boldsymbol{x}|\omega_j, \chi_k(\boldsymbol{x}) = 1)$ in Lemma 1 it follows that

$$P(\boldsymbol{x}|\omega_j, \chi_k(\boldsymbol{x}) = 1) = \begin{cases} P(\boldsymbol{x}|\omega_j), & \text{if } \boldsymbol{x} \in \chi_{\beta(j)} \\ 0, & \text{otherwise.}\end{cases}$$

and, using this result in (26) and (27), that

$$\hat{\boldsymbol{\mu}}_{q,j,k} = \delta_{k,\beta(j)}\boldsymbol{\mu}_{q,j} \tag{37}$$
$$\hat{\boldsymbol{\Sigma}}_{q,j,k} = \delta_{k,\beta(j)}\boldsymbol{\Sigma}_{q,j}. \tag{38}$$

Combining (36), (37), and (38) with (35) leads to (30). □

ACKNOWLEDGMENT

The author would like to thank R. Gray for various insightful discussions and pointers to the literature. The author would also like to thank the anonymous reviewers for comments that helped improve the manuscript.

REFERENCES

[1] J. Bach, "The virage image search engine: An open framework for image management," in *Proc. SPIE Storage and Retrieval for Image and Video Databases*, San Jose, CA, 1996.

[2] S. Belongie, J. Malik, and J. Puzicha, "Matching shapes," in *Proc. Int. Conf. Computer Vision*, Vancouver, BC, Canada, 2001.

[3] D. Bertsekas, *Nonlinear Programming*. Cambridge, MA: Athena Scientific, 1995.

[4] J. De Bonet and P. Viola, "Structure driven image database retrieval," in *Proc. Neural Information Processing Systems*, vol. 10, Denver, CO, 1997.

[5] J. De Bonet, P. Viola, and J. Fisher, "Flexible histograms: A multiresolution target discrimination model," in *Proceedings of SPIE*, vol. 3370-12, E. G. Zelnio, Ed., 1998.

[6] 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.*, no. 24, pp. 1026–1038, Aug. 2002.

[7] A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. B-39, 1977.

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

[9] M. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback–Leibler distance," *IEEE Trans. Image Processing*, vol. 11, pp. 146–158, Feb. 2002.

[10] A. Drake, *Fundamentals of Applied Probability Theory*. New York: McGraw-Hill, 1987.

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

[12] J. Foote, M. Cooper, and U. Nam, "Audio retrieval by rhythmic similarity," in *Proc. 3rd Int. Symp. Musical Information Retrieval*, Paris, France, 2002.

[13] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA: Academic, 1990.

[14] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer, 1992.

[15] R. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag, 1988.

[16] R. Gray, D. Neuhoff, and P. Shields, "A generalization of Ornstein's d-bar distance with applications to information theory," *Ann. Probab.*, vol. 3, pp. 315–328, 1975.

[17] R. Gray, J. Young, and A. Aiyer, "Minimum discrimination information clustering: Modeling and quantization with Gauss mixtures," in *Proc. IEEE Int. Conf. Image Processing*, Thesaloniki, Greece, 2001.

[18] R. Gray, A. Gray, G. Rebolledo, and J. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 708–721, Nov. 1981.

[19] 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.

[20] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Spatial color indexing and applications," *Int. J. Computer Vision*, vol. 35, pp. 245–268, Dec. 1999.

[21] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and format frequencies," *Electron. Commun. Japan*, vol. 53-A, pp. 36–43, 1970.

[22] G. Iyengar and A. Lippman, "Clustering images using relative entropy for efficient retrieval," in *Proc. Int. Workshop on Very Low Bitrate Video Coding*, Urbana, IL, 1998.

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

[24] F. Jensen, *An Introduction to Bayesian Networks*. New York: Springer-Verlag, 1996.

[25] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.

[26] M. Kupperman, "Probabilities of hypothesis and information-statistics in sampling from exponential-class populations," *Ann. Math. Statist.*, vol. 29, pp. 571–574, 1958.

[27] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys–Dokl.*, vol. 10, pp. 707–710, 1966.

[28] E. Levina and P. Bickel, "The earth mover's distance is the Mallows distance: Some insights from statistics," in *Proc. Int. Conf. Computer Vision*, vol. 1, Vancouver, BC, Canada, 2001, pp. 251–256.

[29] J. Li and A. Barron, "Mixture density estimation," in *Proc. Neural Information Processing Systems*, Denver, CO, 1999.

[30] 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.

[31] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. IEEE Int. Conf. Multimedia and Expo.*, Tokyo, Japan, 2001.

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

[33] C. Mallows, "A note on asymptotic joint normality," *Ann. Math. Statist.*, vol. 43, pp. 508–515, 1972.

[34] 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.

[35] J. Mao and A. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recogn.*, vol. 25, no. 2, pp. 173–188, 1992.

[36] G. McLean, "Vector quantization for texture classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 23, pp. 637–649, May/June 1993.

[37] H. Neemuchwala, A. Hero, and P. Carson, "Feature coincidence trees for registration of ultrasound breast images," in *Proc. IEEE Int. Conf. Image Processing*, Thessaloniki, Greece, 2001.

[38] 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 and Retrieval for Image and Video Databases*, San Jose, CA, 1993, pp. 173–181.

[39] D. Ornstein, "An application of ergodic theory to probability theory," *Ann. Probab.*, vol. 1, pp. 43–58, 1973.

[40] A. Papoulis, *The Fourier Integral and its Applications*. New York: McGraw-Hill, 1962.

[41] W. Pennebaker and J. Mitchell, *JPEG: Still Image Data Compression Standard*. New York: Van Nostrand Reinhold, 1993.

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

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

[44] R. Picard, T. Kabir, and F. Liu, "Real-time recognition with the entire Brodatz texture database," in *Proc. IEEE Conf. Computer Vision*, New York, 1993.

[45] 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.

[46] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[47] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Proc. Int. Conf. Computer Vision*, Bombay, India, 1998.

[48] 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 Technology*, vol. 8, pp. 644–655, Sept. 1998.

[49] 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.

[50] J. Smith, "Integrated spatial and feature image systems: retrieval, compression and analysis," Ph.D. dissertation, Columbia Univ., New York, 1997.

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

[52] M. Stricker and A. Dimai, "Color indexing with weak spatial constraints," in *SPIE Storage and Retrieval for Image and Video Databases*, vol. 2670, San Jose, CA, 1996, pp. 29–40.

[53] M. Stricker and M. Orengo, "Similarity of color images," in *SPIE Storage and Retrieval for Image and Video Databases*, San Jose, CA, 1995.

[54] M. Stricker and M. Swain, "The capacity of color histogram indexing," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 1994, pp. 704–708.

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

[56] D. Titterington, A. Smith, and U. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[57] H. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968.

[58] K. Valkealahti and E. Oja, "Reduced multidimensional co-occurrence histograms in texture classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 90–94, Jan. 1998.

[59] S. Vallender, "Calculation of the Wasserstein distance between probability distributions on the line," *Theory Probab. Appl.*, vol. 18, pp. 824–827, 1973.

[60] N. Vasconcelos, "Bayesian models for visual information retrieval," Ph.D. dissertation, MIT, Cambridge, MA, 2000.

[61] ——, "A unified view of image similarity," in *Proc. Int. Conf. Pattern Recognition*, Barcelona, Spain, 2000.

[62] N. Vasconcelos and G. Carneiro, "What is the role of independence for visual recognition?," in *Proc. Europ. Conf. Computer Vision*, Copenhagen, Denmark, 2002.

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

[64] N. Vasconcelos and A. Lippman, "A probabilistic architecture for content-based image retrieval," in *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, Hilton Head, NC, 2000.

[65] L. Wasserstein, "Markov processes with countable state space describing large systems of automata," *Probl. Pered. Inform.*, vol. 5, no. 3, pp. 64–73, 1969.