Appears in International Conference on Artificial Neural Networks, Istanbul, Turkey, 2003.

The design of end-to-end optimal image retrieval systems

Nuno Vasconcelos
Electrical and Computer Engineering,
University of California, San Diego,
nuno@ece.ucsd.edu

***Abstract*— In the last few years, image retrieval has evolved from early simplistic solutions (such as histogram intersection) to more principled approaches that can be shown to be optimal under sensible criteria for the retrieval problem. We discuss a decision theoretic formulation that enables the design of systems where all components are optimized with respect to the same end-to-end performance criteria: the minimization of the probability of retrieval error. This discussion includes both a theoretical characterization of how the probability of error is affected by the design of the basic components of a retrieval system (feature transformation, feature representation, and similarity function) and experimental evidence of how various design choices impact the retrieval performance.**

I. Introduction

An architecture for content-based image retrieval (CBIR) consists of three fundamental building blocks: 1) a feature transformation from the space of image observations (e.g. pixels) to a feature space with better retrieval properties, 2) a feature representation that compactly describes how each of the database image classes populates this space, and 3) a similarity function that allows ranking the database classes by similarity to a query. Ideally, one would like to design retrieval systems where all retrieval components are optimized with respect to the same performance criteria. Since the ultimate goal of any retrieval system is to be correct as often as possible, a natural choice is the *minimization of probability of retrieval error* (MPE). It leads to a *generic* retrieval architecture that is not tied to a particular type of imagery or database, and it makes a vast body of existing decision-theoretic results applicable to the retrieval problem. In this paper we summarize some of our recent work on MPE retrieval systems [1]–[4].

II. Decision-theoretic image similarity

A retrieval system is a mapping from a feature space $\mathcal{X}$ to the index set of the M classes in the database.

Definition 1: Given a feature space $\mathcal{X}$ and a set of M image classes Y, a minimum probability of error (MPE) retrieval system is the mapping

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

that minimizes

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

Under this definition, the optimal similarity function is automatically determined by the following well known theorem [5].

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

$$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}). \tag{1}$$

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

$$L^*_{\mathcal{X}} = 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \tag{2}$$

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

Proof: See [7] for all proofs. ■

The MAP classifier can be implemented by application of Bayes rule

$$g^*(\mathbf{x}) = \arg\max_i \sum_{j=1}^{N} \log P_{\mathbf{X}|Y}(\mathbf{x}_j|i) + \log P_Y(i), \tag{3}$$

where $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ is the class-conditional likelihood for the i^{th} class, $P_Y(i)$ a *prior probability* for this class, and we have used assumed that the feature vectors in $\mathbf{x}$ are mutually independent. Equation (3) is denoted by *Bayesian retrieval criteria* and image retrieval based on it as *decision-theoretic retrieval* (DTR). It requires the selection of a feature space and a feature representation.

III. Decision-theoretic guidelines for image representation

The following theorem shows that the choice of features determines the lowest possible error that can be achieved for a given database.

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

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

then

$$L^*_{\mathcal{X}} \geq L^*_{\mathcal{Z}} \tag{4}$$

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

The theorem shows that, by introducing a feature transformation, it is never possible to decrease the Bayes error. On the contrary, this lower bound on the error probability will increase in all cases except where the feature transformation is invertible. While a necessary condition, low Bayes error is not sufficient to guarantee accurate retrieval since the actual error may be much larger than the lower bound. In the remainder of this work we assume that the classes are a-priori equiprobable, i.e. $P_Y(i) = 1/M, \forall i$. This leads to the following theorem.

Theorem 3: Given a retrieval problem with equiprobable classes, a feature space $\mathcal{X}$, unknown class conditional likelihood functions $P_{\mathbf{X}|Y}(\mathbf{x}|i)$, and a decision function

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

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

$$P(g(\mathbf{X}) \neq Y) - L^*_{\mathcal{X}} \leq \Delta_{g,\mathcal{X}} \tag{6}$$

where

$$\Delta_{g,\mathcal{X}} = \sum_i KL[P_{\mathbf{X}|Y}(\mathbf{x}|i)||\hat{p}_{\mathbf{X}|Y}(\mathbf{x}|i)], \tag{7}$$

is the estimation error.

The theorem shows that the difference between the actual probability of retrieval error and the Bayes error is upper bounded by the error in the density estimates. It follows that, if the Bayes error is small, accurate density estimation is a sufficient condition for high retrieval accuracy. This observation has an immediate impact on the selection of the probabilistic model, or feature representation, used to estimate the unknown densities $P_{\mathbf{X}|Y}(\mathbf{x}|i)$: it should be flexible enough to enable accurate density estimates in high-dimensional spaces. In this context, we have compared the properties of various probabilistic models and shown that the Gaussian mixture models exhibit various attractive properties.

Definition 2: A Gaussian mixture is a density of the form

$$P_{\mathbf{X}}(\mathbf{x}) = \sum_{w=1}^{C} \pi_w \mathcal{G}(\mathbf{x}, \mu_w, \mathbf{\Sigma}_w) \tag{8}$$

where

$$\mathcal{G}(\mathbf{x}, \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} e^{-\frac{1}{2}||\mathbf{x}-\mu||^2_{\mathbf{\Sigma}}} \tag{9}$$

is the Gaussian density of mean μ and covariance $\mathbf{\Sigma}$, and

$$||\mathbf{x} - \mu||_{\mathbf{\Sigma}} = \sqrt{(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)}. \tag{10}$$

See [3], [4] for more details.

IV. Optimal feature transformations

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

A. Embedded feature spaces

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

$$\epsilon : \mathcal{X}_m \to \mathcal{X}_n \tag{11}$$

which is one-to-one.

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

$$\iota_m^n : \mathbb{R}^m \to \mathbb{R}^n \tag{12}$$

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

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

$$\epsilon_i : \mathcal{X}_i \to \mathcal{X}'_{i+1}, \; i = 1, \ldots, d-1, \tag{13}$$

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

The inverse operation of an embedding is a submersion.

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

$$\gamma : \mathcal{X}_n \to \mathcal{X}_m \tag{14}$$

which is surjective.

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

$$\pi_m^n : \mathbb{R}^n \to \mathbb{R}^m \tag{15}$$

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

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

Theorem 4: Let

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

be a linear feature transformation. Then,

$$\mathcal{X}_i = \pi_i^d(\mathcal{X}), i = 1, \ldots, d-1 \tag{16}$$

is a sequence of embedded feature spaces such that

$$L^*_{\mathcal{X}_{i+1}} \leq L^*_{\mathcal{X}_i}. \tag{17}$$

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

$$\mathbf{X}_i = \pi_i^d(\mathbf{X}), i = 1, \ldots, d \tag{18}$$

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

$$g_i(\mathbf{x}) = \arg\max_k \hat{p}_{\mathbf{X}_i|Y}(\mathbf{x}|k) \tag{19}$$

then

$$\Delta_{g_{i+1}, \mathcal{X}_{i+1}} \geq \Delta_{g_i, \mathcal{X}_i}. \tag{20}$$

It follows that, in general, it is impossible to minimize the Bayes and estimation errors simultaneously. On one hand, given a feature space $\mathcal{X}_i$ it is usually possible to find a subspace where density estimates are more accurate. On the other, the projection onto this subspace will increase the Bayes error. The practical result is that there is always a need to reach a compromise between the two sources of error. This is illustrated by Figure 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 $\mathcal{X}$.

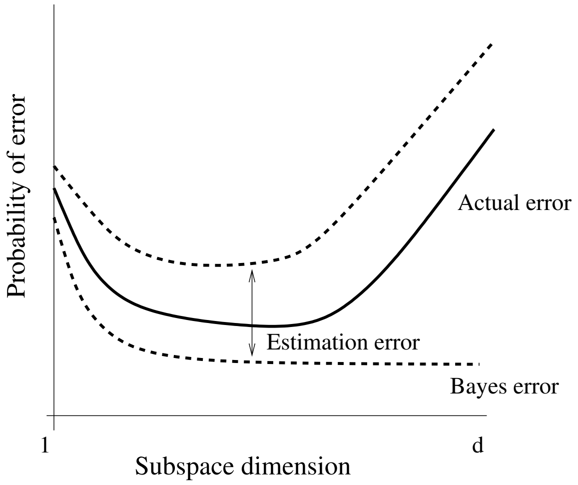

Fig. 1. Upper bound, lower bound, and probability of error as a function of subspace dimension.

Since accurate density estimates can usually be obtained in low-dimensional spaces, the two bounds tend to be close when the subspace dimension is small. In this case, the probability of error is dominated by the Bayes error. For higher-dimensional subspaces the decrease in Bayes error is canceled by an increase in estimation error and the actual probability of error increases. Overall, the curve of the

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

- *for each query image $I_s \in I$:*
  - *apply the transformation T to a collection of observations from I_s to obtain a set of query feature vectors $\mathbf{x}_s = \{\mathbf{x}_{s,1}, \ldots, \mathbf{x}_{s,N}\}$*
  - *for each subspace dimension $j = 1, \ldots, d$*
    - *for each image class $i = 1, \ldots, M$*
      - *apply (22) to obtain the embedded mixtures $P_{\mathbf{X}_j|Y}(\mathbf{x}|i)$*
      - *compute*

$$p^i_{s,j} = \frac{1}{N}\sum_{k=1}^{N} \log P_{\mathbf{X}_j|Y}(\pi^d_j(\mathbf{x}_{s,k})|i).$$

    - *sort the $p^i_{s,j}$ by decreasing value and, based on the resulting order, evaluate some measure of retrieval performance (e.g. precision at some level of recall) $R_{s,j}$.*
- *average the retrieval measure across queries $R_j = 1/|I| \sum_s R_{s,j}$.*
- *return the subspace dimension $j^* = \arg\max_j R_j$ and associated performance score R_{j^*}.*

Fig. 2. Algorithm for determining the optimal subspace dimension for a retrieval problem with feature transformation T, and class densities $\{P_{\mathbf{X}|Y}(\mathbf{x}|i), i = 1, \ldots, M\}$.

probability of error exhibits the convex shape depicted in the figure, where an inflection point marks the subspace dimension for which Bayes error ceases to be dominant. Hence, given a feature space $\mathcal{X}$, the *MPE-optimal retrieval system is the one which operates on this inflection point*.

The next lemma shows that, under the mixture representation, it is trivial to derive the density estimates on each of the embedded subspaces $\mathcal{X}_j$ from the density on $\mathcal{X}$.

Lemma 1: Let $\mathcal{X}$ be a feature space, $\{\mathcal{X}_j\}$ a sequence of embedded subspaces according to (16), and $\mathbf{X}^d_1$ a sequence of random vectors according to (18). If, under class i, $\mathbf{X}$ is distributed according to the Gaussian mixture density

$$P_{\mathbf{X}|Y}(\mathbf{x}|i) = \sum_{c=1}^{C} \pi_{i,c}\mathcal{G}(\mathbf{x}, \mu_{i,c}, \mathbf{\Sigma}_{i,c}) \tag{21}$$

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

$$P_{\mathbf{X}_j|Y}(\mathbf{x}|i) = \sum_{c=1}^{C} \pi_{i,c}\mathcal{G}[\mathbf{x}, \mathbf{\Pi}^d_j\mu_{i,c}, \mathbf{\Pi}^d_j\mathbf{\Sigma}_{i,c}(\mathbf{\Pi}^d_j)^T], \tag{22}$$

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

The collection of densities in (22) is denoted by the family of *embedded mixture models* (EMMs) associated with $\mathbf{X}$. Notice that once an estimate is available for $\{\pi_{i,c}, \mu_{i,c}, \mathbf{\Sigma}_{i,c}\}$ the parameters of $P_{\mathbf{X}_j|Y}(\mathbf{x}|i)$ are obtained by simply extracting the first j components of the mean vectors $\mu_{i,c}$ and the upper-left $j \times j$ sub-matrix of the covariances $\mathbf{\Sigma}_{i,c}$. The lemma suggests an efficient cross-validation procedure to find the optimal subspace dimension of a given transformation T: select a set of query images $I = \{I_1, \ldots, I_Q\}$, establish the associated retrieval ground truth, and use this set to infer the optimal subspace dimension. An algorithmic description of this procedure is given in Figure 2. It remains to determine how the feature transformation T can itself be found. One possibility, that we explore next, is to restrict the search to a finite dictionary of transformations that satisfy some properties known to be important for visual recognition, e.g. invariance to certain image mappings or plausibility under what is known about human perception.

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

1) *select a reference transformation in $\mathcal{T}$, e.g. $T^{(1)}$;*
2) *for each image class $i = 1, \ldots, M$, use a standard maximum likelihood estimation technique, e.g. the expectation-maximization algorithm [6], to determine the mixture parameters of $P_{\mathbf{X}^{(1)}|Y}(\mathbf{x}|i)$;*
3) *for each transformation $f = 2, \ldots, F$*
   - *let $T^{(1,f)} = T^{(f)} \circ (T^{(1)})^{-1}$*
   - *compute, for each image class $i = 1, \ldots, M$, the parameters of $P_{\mathbf{X}^{(f)}|Y}(\mathbf{x}|i)$ using (25) and (26).*
   - *let $(j^*_f, R^*_f) =$ **subspacedim**$(I, T^{(f)}, \{P_{\mathbf{X}^{(f)}|Y}(\mathbf{x}|i), i = 1, \ldots, M\})$*
4) *let $f^* = \arg\max_f R^*_f$ and $j^* = j^*_{f^*}$;*
5) *return $T^{(f^*)}_{j^*} = \pi^d_{j^*}(T^{(f^*)})$*

Fig. 3. Algorithm for determining the best feature transformation, and subspace dimension for a retrieval problem with transformation dictionary $\mathcal{T}$.

B. Optimal features

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

$$P_{\mathbf{X}^{(l)}|Y}(\mathbf{x}|i) = \sum_{c=1}^{C} \pi_{i,c}\mathcal{G}(\mathbf{x}, \mu^l_{i,c}, \mathbf{\Sigma}^l_{i,c}) \tag{23}$$

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

$$P_{\mathbf{X}^{(m)}|Y}(\mathbf{x}|i) = \sum_{c=1}^{C} \pi_{i,c}\mathcal{G}(\mathbf{x}, \mu^m_{i,c}, \mathbf{\Sigma}^m_{i,c}) \tag{24}$$

where

$$\mu^m_{i,c} = \mathbf{T}^{(l,m)}\mu^l_{i,c} \tag{25}$$

$$\mathbf{\Sigma}^m_{i,c} = \mathbf{T}^{(l,m)}\mathbf{\Sigma}^l_{i,c}(\mathbf{T}^{(l,m)})^T . \tag{26}$$

Therefore, it suffices to perform density estimation on a reference subspace, e.g. $\mathcal{X}^{(1)}$, in order to obtain the mixture parameters associated with all transformations in $\mathcal{T}$. The search for the optimal feature transformation can thus be performed with the algorithm of Figure 3.

V. Experimental evaluation

For a number of years we have been performing extensive evaluation of the impact, on the probability of error, of various aspects of the three components of a retrieval system [1], [4], [7], [8]. Here, we will only offer a small subset of these results, oriented to illustrate how sub-optimal design decisions can have a significant influence on the overall retrieval performance. We present results on the widely used Brodatz (1008 texture images) and Corel (1500 color images

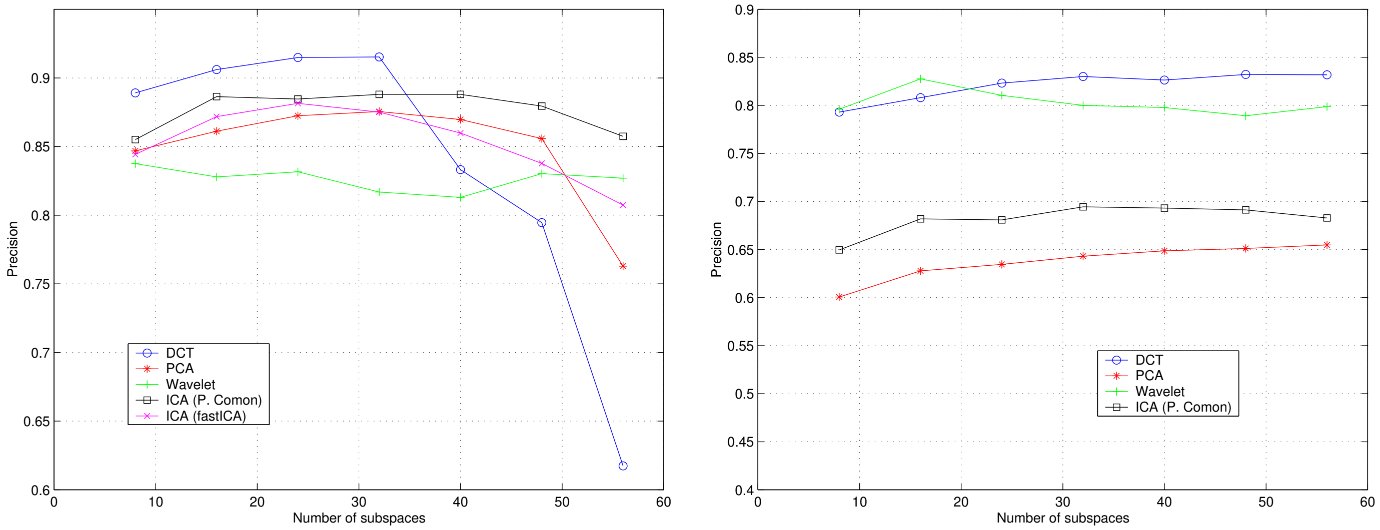

Fig. 4. Left: Precision, at 30% recall, on Brodatz. Right: Precision, at 30% recall, on Corel.

of natural scenes) databases. All mixtures contain 8 Gaussians of diagonal covariance, and all mixture parameters were learned with the EM algorithm [6]. Each image in the database was considered as a different class, and a uniform prior was assumed for the image classes.

This set of experiments was designed to assess the retrieval performance of various feature transformations. For reasons that are detailed in [7] we considered five multi-resolution transforms: discrete cosine transform (DCT), a wavelet decomposition (Daubechies'), principal component analysis (PCA), and two independent component analysis techniques (ICA). In all cases, the region of support of the features was such to originate 64 coefficients per color channel, which were interleaved according to the pattern YBRYBR... Figure 4 presents curves of precision, as a function of subspace dimension, at 30% recall on the two databases.

Since precision is inversely proportional to the probability of error one would expect, from the theoretical arguments of section IV-A, the precision curves to be concave. This is indeed the case (there is a large increase in precision from 1 to 8 dimensions that we do not show for clarity of the graph) for all transformations. In terms of the relative performance of the different transforms, the DCT is the top performer for both databases reaching high precision in both cases. On the other hand, PCA always performs poorly. This is a interesting result, given that PCA has been widely used in visual recognition [9], [10]. The performance of the other features seems to be significantly more dependent on the database. Wavelets do quite well on Corel, but very poorly on Brodatz, ICA does better on Brodatz than on Corel.

The main insight to be retained from these experiments is that a careless choice of the feature transformation can lead to very poor retrieval performance. On Brodatz the peak precision of the worst transformation (wavelet) is 10% below that of the best (DCT), on Corel the variation is almost 20%. Even for a given transformation, precision can vary dramatically with the number of embedded subspaces. For example, the precision of the DCT features on Brodatz drops from the peak value of about 92% to about 62% when all the subspaces are included. These observations emphasize the importance of the feature selection algorithm discussed in section IV-B.

REFERENCES

[1] N. Vasconcelos and A. Lippman, "A Probabilistic Architecture for Content-based Image Retrieval," in *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, Hilton Head, North Carolina, 2000.

[2] N. Vasconcelos, "On the Complexity of Probabilistic Image Retrieval," in *Proc. International Conference on Computer Vision*, Vancouver, Canada, 2001.

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

[4] N. Vasconcelos, "Bayesian Models for Visual Information Retrieval," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

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

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

[7] N. Vasconcelos, "Decision-theoretic Image Retrieval with Embedded Multi-resolution Mixtures," Compaq Cambridge Research Laboratory, Tech. Rep. 2002/04, 2002, available from http://crl.research.compaq.com.

[8] ——, "Decision-theoretic Image Retrieval," in *SPIE Multimedia Management Systems III*, Boston, 2002.

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

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