Appears in Proc. IEEE Conference Computer Vision and Pattern Recognition,
Hilton Head, North Carolina, 2000.

A Probabilistic Architecture for Content-based Image Retrieval

Nuno Vasconcelos and Andrew Lippman
MIT Media Lab, www.media.mit.edu/~{nuno,lip}

Abstract

The design of an effective architecture for content-based retrieval from visual libraries requires careful consideration of the interplay between feature selection, feature representation, and similarity metric. We present a solution where all the modules strive to optimize the same performance criteria: the probability of retrieval error. This solution consists of a Bayesian retrieval criteria (shown to generalize the most prevalent similarity metrics in current use) and an embedded mixture representation over a multiresolution feature space (shown to provide a good trade-off between retrieval accuracy, invariance, perceptual relevance of similarity judgments, and complexity). The new representation extends standard models (histogram and Gaussian) by providing simultaneous support for high-dimensional features and multi-modal densities and performs well on color, texture, and generic image databases.

1 Introduction

An architecture for image retrieval is composed by three fundamental building blocks: a feature transformation, a feature representation and a similarity metric. Even though significant attention has been recently devoted to each of these individual components, there have been significantly less attempts to investigate the interrelationships between them and how these relationships may affect the overall performance of retrieval systems.

Current retrieval solutions can be grouped into two major disjoint sets: the ones tailored for texture vs the ones tailored for color. These two sets vary widely with respect to the emphasis placed on the design of the individual retrieval components. Because most texture databases consist of homogeneous images, texture retrieval usually assumes Gaussian distributed features for which simple similarity metrics, such as the Euclidean or Mahalanobis distances, are optimal. The focus is, instead, on finding the feature transformation that leads to best discrimination between the different texture classes.

On the other hand, feature selection has not been a critical issue for color-based retrieval, where the features are usually the pixel colors themselves. Instead a significant amount of work has been devoted to the issue of feature representation, where the majority of the proposed solutions are variations on the color histogram initially proposed for object recognition [9], e.g. color coherence vectors, color correlograms, color moments, etc. The most common similarity metrics are L^p norms and, among these, the L^1 distance (histogram intersection [9]) has become quite popular.

While they have worked well in their specific domains, these representations break down when applied to databases of generic imagery. The main problem for texture-based solutions is that, since generic images are not homogeneous, their features cannot be accurately modeled as Gaussian and simple similarity metrics are no longer sufficient. On the other hand, color-based solutions are plagued by the exponential complexity of the histogram on the dimension of the feature space, and are applicable only to low-dimensional features (e.g. pixel colors). Hence, they are unable to capture the spatial dependencies that are crucial for texture characterization.

In the absence of solutions that can account for both color and texture, retrieval systems must resort to different features, representations and similarity functions to deal with the two image attributes [3], making it difficult to perform joint inferences with respect to both. The standard solution is to evaluate similarity according to each of the attributes and obtain an overall measure by weighting linearly the individual distances. This opens up the question of how to weigh different representations on different feature spaces, a problem that has no easy solution.

In order to overcome these problems we present an integrated solution for image retrieval, where all three modules are designed with respect to the same performance criteria: *minimization of the probability of retrieval error*. This is shown to be interesting in two ways. First, it leads to a Bayesian formulation of retrieval and a probabilistic retrieval criteria that either generalizes or improves upon the most commonly used similarity functions (Mahalanobis distance, L^p norms,and Kullback-Leibler divergence). Second, it shows that the most restrictive constraints on the design of the retrieval architecture are actually imposed on feature selection. In fact, optimal performance can only be achieved under a restricted set of *invertible transformations* that leaves small space for feature optimization.

A corollary, of great practical relevance, of these two observations is that for retrieval from generic databases there is less to be gained from feature selection than from a carefully designed feature representation. In this context, being expressive enough to capture the details of multimodal densities and compact enough to be tractable in high-dimensions, the Gaussian mixture emerges as the right generalization, for joint modeling of color and texture, to the standard histogram and Gaussian models.

Further observation that 1) a mixture model defines a family of embedded densities, and 2) the linear invertible feature transformation that provides optimal compaction (principal component analysis) is well approximated by (perceptually more justifiable) multiresolution transformations such as the DCT leads to the concept of embedded multiresolution mixtures (EMM). These are a family of embedded densities ranging over multiple image scales that allow explicit control over the trade-off between spatial support and invariance.

Overall the new retrieval architecture provides a good trade-off between retrieval accuracy, invariance, perceptual relevance of similarity judgments, and complexity. This conclusion is supported by detailed experimental evaluation showing that the new solution outperforms the previous best for both objective (precision/recall) and subjective (perceptual) evaluations.

2 Probabilistic retrieval

The problem of retrieving images or video from a database is naturally formulated as a problem of classification. Given a representation (or feature) space $\mathcal{F}$ for the entries in the database, the design of a retrieval system consists of finding a map

$$\begin{array}{rccl} g: & \mathcal{F} & \rightarrow & M = \{1, \ldots, K\} \\ & \mathbf{x} & \rightarrow & y \end{array}$$

from $\mathcal{F}$ to the set M of classes identified as useful for the retrieval operation.

In this work, we define the goal of a content-based retrieval system to be the *minimization of the probability of retrieval error*, i.e. the probability $P(g(\mathbf{x}) \neq y)$ that if the user provides the retrieval system with a set of feature vectors $\mathbf{x}$ drawn from class y the system will return images from a class $g(\mathbf{x})$ different than y. Once the problem is formulated in this way, it is well known that the optimal map is the Bayes classifier [1]

$$\begin{aligned} g^*(\mathbf{x}) &= \arg\max_i P(y = i|\mathbf{x}) && (1) \\ &= \arg\max_i P(\mathbf{x}|y = i) P(y = i), && (2) \end{aligned}$$

where $P(\mathbf{x}|y = i)$ is the likelihood function, or feature representation, of the features from the i^{th} class and $P(y = i)$ the prior probability for this class. When all classes are a priori equally likely this leads to the standard maximum-likelihood (ML) classifier

$$g(\mathbf{x}) = \arg\max_i P(\mathbf{x}|y = i) \quad (3)$$

or, if instead of a single feature vector $\mathbf{x}$ we have a collection of N independent query features, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

$$g(\mathbf{X}) = \arg\max_i \frac{1}{N} \sum_{j=1}^{N} \log P(\mathbf{x}_j|y = i). \quad (4)$$

Some of the most popular retrieval criteria in current use are special cases of this probabilistic formulation. For example, when N is large, the simple application of the law of large numbers reveals that

$$\begin{aligned} g(\mathbf{X}) &\rightarrow \arg\max_i E_{\mathbf{x}}^q[\log P(\mathbf{x}|y = i)] \\ &= \arg\max_i \int P(\mathbf{x}|q) \log P(\mathbf{x}|y = i) dx \quad (5) \\ &= \arg\min_i \int P(\mathbf{x}|q) \log P(\mathbf{x}|q) dx \\ &\quad - \int P(\mathbf{x}|q) \log P(\mathbf{x}|y = i) dx \\ &= \arg\min_i KL(Q||P_i) \end{aligned}$$

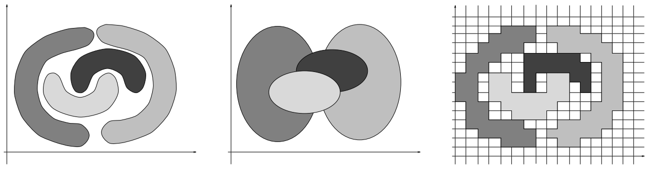
where $P(\mathbf{x}|q)$ is the density of the query features, and $KL(Q||P_i)$ the Kullback-Leibler divergence between this density and that associated with the i^{th} database image class [1]. I.e. the KL divergence is the asymptotic limit of the ML criteria. While the two criteria perform equally well for *global queries* based on entire images, ML has the added advantage of also enabling *local queries* consisting of much smaller user-selected image regions.

When all the likelihoods are assumed Gaussian, equation (5) reduces to

$$g(\mathbf{X}) = \arg\min_i \log|\mathbf{\Sigma}_i| + trace[\mathbf{\Sigma}_i^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}}] + (\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\hat{\mathbf{x}} - \mu_i)^T \quad (6)$$

where $\hat{\mathbf{x}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{x}}$ are the sample mean and covariance of $\mathbf{X}$ and μ_i and $\mathbf{\Sigma}_i$ the mean and covariance of $P(\mathbf{x}|y = i)$ [10]. The third term in this equation is the widely used *Mahalanobis distance* and the two other terms can be shown to augment this distance, enabling it to detect changes in scale and orientation of the query density. This can lead to significant improvements in retrieval accuracy [10, 11].

Finally, because the probability of error of (4) tends to the probability of error of the Bayes classifier orders of magnitude faster than the associated density estimates tend to the right distributions [1], the ML criteria places much



a) b) c)

Figure 1. a) four image classes, b) Gaussian fits, c) histogram fits.

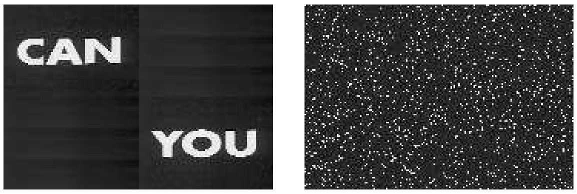


Figure 2. Two visually distinct images that have the exact same color histogram.

less stringent requirements on the quality of these estimates than criteria based on the minimization of L^p distances, such as the *histogram intersection* widely used for color retrieval. Since density estimation is a difficult problem this can lead to significant improvements also with respect to this class of techniques [11].

3 Trading Bayes error for complexity

In addition to the ML retrieval criteria, the goal of minimizing retrieval error also provides guidance for feature selection. In particular, one would like to get as close as possible to the least possible error achievable with the Bayes classifier of (2), the Bayes error [1]

$$L^* = 1 - E_{\mathbf{x}}[\max_i P(y = i|\mathbf{x})].$$

Whenever a feature transformation $T : R^n \rightarrow R^k$ is employed the Bayes error becomes L', where

$$\begin{aligned} 1 - L' &= E_{T(\mathbf{x})}[\max_i P(y = i|T(\mathbf{x}))], \\ &= E_{T(\mathbf{x})}[\max_i E_{\mathbf{x}|T(\mathbf{x})}[P(y = i|\mathbf{x})|T(\mathbf{x}))], \\ &\leq E_{\mathbf{x}}[\max_i P(y = i|\mathbf{x})] = 1 - L^*. \end{aligned}$$

I.e., the choice of feature space has a direct impact on this optimal performance bound: 1) any feature transformation can only increase or, at best, maintain the Bayes error achievable in the original space of image pixels, and 2) the only transformations which maintain the Bayes error are the invertible ones. This indicates that feature sets which arbitrarily discard information are a bad idea.

Because discarding information reduces complexity and increases invariance to image transformations, the interesting question is then how to discard dimensions in a way that compromises as little as possible the achievable Bayes error. We rely on the strategy of discarding those features whose deletion leads to a transformation that is as close to invertible as possible. If $\mathbf{x} \in R^n$, $\mathcal{T} : R^n \rightarrow R^n$ is an invertible transformation and $\mathcal{Q}_i : R^n \rightarrow R^n$ the map $\mathcal{Q}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$ we successively look for

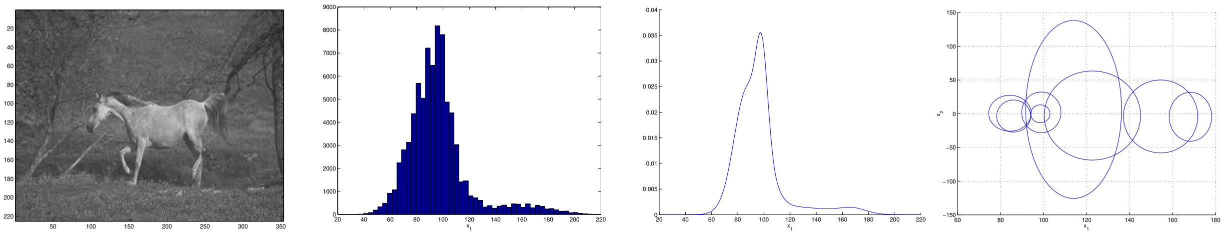
$$j = \arg\min_{i \in I} E[||\mathcal{T}^{-1}(\mathcal{Q}_i(\mathcal{T}(\mathbf{x}))) - \mathbf{x}||]$$

where I is the set of feature indices. When $\mathcal{T}$ is restricted to be a linear transformation, the optimal solution is provided by the well known PCA dimensionality reduction technique. It is also well known that, for local image neighborhoods, PCA is well approximated by frequency decompositions, such as the DCT, that are simpler to compute and better matched to human perception. This makes the space of DCT coefficients a natural feature space for CBIR from the Bayes error, perceptual, and complexity points of view.

4 Feature representation

Maintaining the Bayes error small is, in practice, not sufficient to guarantee good retrieval accuracy, since the degree to which this lower bound can be attained is a function of quality of the estimates of $P(\mathbf{x}_j|y = i)$ in (4). Unfortunately, there are serious limitations associated with the feature representations in common use in the retrieval literature. These are illustrated in Figure 1 where we depict an hypothetical two-dimensional retrieval problem with four image classes. The class densities are represented in figure a) by the contour where the probability drops to, say, 50% of its maximum value. The best Gaussian fit to each of the class densities under the Gaussian model, implicit in the use of the Mahalanobis distance, is shown in b). Because the Gaussian model is too simplistic to capture the details of the true densities there is a lot of class-overlap and the classification error is high.

As illustrated by Figure c) a significantly better approximation can be achieved with the histogram model, widely used for color retrieval. However, this comes at a price: an exponential dependence of the model complexity (number of histogram bins) in the dimension of the feature space, that prohibits the use of histograms in high dimensional spaces. The ability to cope with high-dimensional features is a requirement when one wants to deal with spatially supported features that can capture spatial image dependencies.



a) b) c) d)

Figure 3. a) an image from the Corel database, b) its histogram, c) projection of the corresponding 64 dimensional embedded multiresolution mixture onto the DC subspace, and d) projection onto the subspace of two lower frequency coefficients (contours where likelihood drops to 60%).

Such dependencies are an essential component of image properties like texture or local surface appearance and, as shown in Figure 2, crucial for fine image discrimination.

An alternative feature representation is the Gaussian mixture model

$$P(\mathbf{x}|y=i) = \sum_{c=1}^{C} \pi_i \mathcal{G}(\mathbf{x}, \mu_c, \mathbf{\Sigma}_c), \quad (7)$$

where $\mathcal{G}(\mathbf{x}, \mu_c, \mathbf{\Sigma}_c)$ is a Gassian of mean μ and covariance $\mathbf{\Sigma}$. Gaussian mixtures are 1) able to approximate arbitrary densities and 2) computationally tractable on high dimensions (complexity only quadratic in the dimension of the feature space).

5 Embedded multiresolution mixtures

The simple adoption of a mixture model on a high dimensional space does not automatically solve all the representation problems. In particular, it is well know that, as the region of support of the features increases it is also increasingly more difficult to make the representation invariant to image transformations. There are several ways in which invariance can be achieved: filtering out high-frequency information [7] encoding invariance into the similarity function [8], or simply including examples covering all types of variation in the training set [6].

From the Bayes error point of view, the two latter solutions are preferable because they imply discarding no information. Furthermore, since explicit modeling of all transformations in the similarity function significantly increases the retrieval complexity, learning is the best solution. Nevertheless, due to its combinatorial complexity in the number of degrees of freedom to be accounted for, it is usually impossible to rely on learning uniquely. In practice, it is instead necessary to reach a balance between explicit encoding of invariance in the features and learning invariance through training. This leads to the idea of EMM models.

The key observation is that the restriction of a Gaussian in $\mathcal{R}^n$ to $\mathcal{R}^k$, $k \leq n$ is still a Gaussian. In particular, if $\mathbf{x} \in \mathcal{R}^n$ and $P(\mathbf{x}) = \mathcal{G}(\mathbf{x}, \mu_x, \mathbf{\Sigma}_x)$ then

$$P(\mathbf{x})|_{\mathcal{R}^k} = P(\mathbf{\Gamma}_k \mathbf{x}) = \mathcal{G}(\mathbf{\Gamma}_k \mathbf{x}, \mathbf{\Gamma}_k \mu_x, \mathbf{\Gamma}_k \mathbf{\Sigma}_x \mathbf{\Gamma}_k^T) \quad (8)$$

where $\mathbf{\Gamma}_k = [\mathbf{I}_k \, \mathbf{0}_{n-k}]$, $\mathbf{I}_k$ ($\mathbf{0}_{n-k}$) is the identity (zero) matrix of order k ($n-k$) and the result follows from the fact that the projection into $\mathcal{R}^k$ is a linear transformation. Combining equations (7) and (8) we obtain a similar result for Gaussian mixtures

$$P(\mathbf{x}|y=i)|_{\mathcal{R}^k} = \sum_{c=1}^{C} \pi_c \mathcal{G}(\mathbf{\Gamma}_k \mathbf{x}, \mathbf{\Gamma}_k \mu_c, \mathbf{\Gamma}_k \mathbf{\Sigma}_c \mathbf{\Gamma}_k^T). \quad (9)$$

This implies that the set of parameters $\{\pi_c, \mu_c, \mathbf{\Sigma}_c\}$ defines a family of embedded densities $\{P(\mathbf{x}|y=i)|_{\mathcal{R}^k}\}_{k=1}^n$.

When, as is the case of the DCT features, the underlying feature space results from a multiresolution decomposition this leads to an interesting interpretation of the mixture density as a family of densities defined over multiple image scales, each adding higher resolution information to the characterization provided by those before it. Disregarding dimensions associated with high frequency basis functions is therefore equivalent to modeling densities of low-pass filtered images. In the extreme case where only the first, or DC, coefficient is considered the representation is equivalent to the histogram. This is illustrated in Figure 3.

The EMM model can thus be seen as a generalization of the color histogram, where the additional dimensions capture the spatial dependencies that are crucial for fine image discrimination (as illustrated in Figure 2). This generalization also enables fine control over the invariance properties of the representation. Since the histogram is approximately invariant to scaling, rotation and translation, when only the DC subspace is considered the representation is invariant to these transformations. As high-frequency coefficients are included invariance is gradually sacrificed. Of course, invariance can always be improved by including proper examples in the training sample.

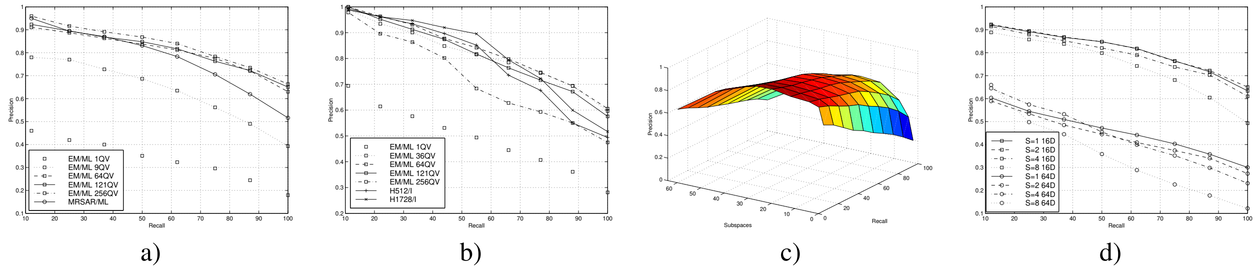


Figure 4. Precision/recall curves on a) Brodatz and b) Columbia. For comparison, curves obtained with MRSAR/ML and HI are also shown. X QV means that only X feature vectors were included in the query. c) surface spanned by precision/recall as a function of the number of subspaces considered, for Brodatz. d) Precision/recall as a function of block spacing (S) during learning when 16 and 64 subspaces are considered.

6 Experimental evaluation

To compare the performance of probabilistic retrieval with that of standard solutions, we conducted experiments on two different databases: the Brodatz texture database, and the Columbia object database. While Brodatz provides a good testing ground for texture retrieval, color-based methods tend to do well on Columbia.

The 1008 images in the Brodatz database were divided into two subgroups: a *query database* of 112, and a *retrieval database* of 896 images. The retrieval performance of the combination of the MRSAR features and Mahalanobis distance (MD) (following the implementation in [5]) was used as a benchmark for this database. The Columbia database was also split into two subsets: a query database containing a single view of each of the 100 objects available, and a retrieval database containing 9 views (separated by 40^o) of each object. Histogram intersection (HI) [9] was used as a benchmark on Columbia.

All images were converted from the original RGB to the YBR color space. Unless otherwise noted, DCT features were obtained with an 8×8 window sliding by increments of two pixels. Mixtures of 8 Gaussians were used for the Brodatz and Corel databases and 16 for Columbia. Only the first 16 embedded subspaces (DCT coefficients) were considered for retrieval. Each image in the database was considered as an independent class.

6.1 Embedded mixtures

Figures 4 a) and b) present precision/recall curves of ML retrieval with EMM for the Brodatz and Columbia databases against those of MRSAR/ML (Brodatz) and HI (Columbia). Because ML can be used both as a measure of local or global similarity, we performed a series of experiments where the query consisted of only a few of the feature vectors available in the query image. From a total of 256 non-overlapping blocks, the number used in each experiment varied between 1 (0.3% of the image) and 256 (100%)[1]. Blocks were selected starting from the center in an outward spiral fashion.

Two conclusions can be drawn from the figure. First, when used as a global similarity metric (a significant portion of the feature vectors in the query image selected) the EMM/ML combination achieves equivalent performance or actually outperforms the retrieval approaches that are specific for the domain of each database (MRSAR/ML on Brodatz and HI on Columbia). This shows that EMM/ML can handle both color and texture indicating that the representation should do well across a large spectrum of databases. Second, a small subset of the query feature vectors is sufficient to achieve retrieval performance close to the best. In both cases 64 query vectors, 0.4% of the total of features that could be extracted from the image and covering only 25% of its area, are enough. In summary, ML has good properties both as a local and a global metric of similarity and is very robust against missing data.

From a perceptual standpoint, the results achieved with EMM/ML are also superior to those obtained with the MRSAR and histogram-based approaches. In particular, EMM/ML has three major advantages: 1) when it commits errors, these errors tend to be perceptually less annoying than those originated by the other approaches, 2) when there are several visually similar classes in the database, images from these classes tend to be retrieved together, and 3) even when the performance in terms of precision/recall is worse than that of the other approaches, the results are frequently better form a perceptual point of view. Figure 5 gives examples of each of these types of situations.

[1] Notice that even 256 vectors are a very small percentage (1.5%) of the total number of blocks that could be extracted from the query image if overlapping blocks were allowed.

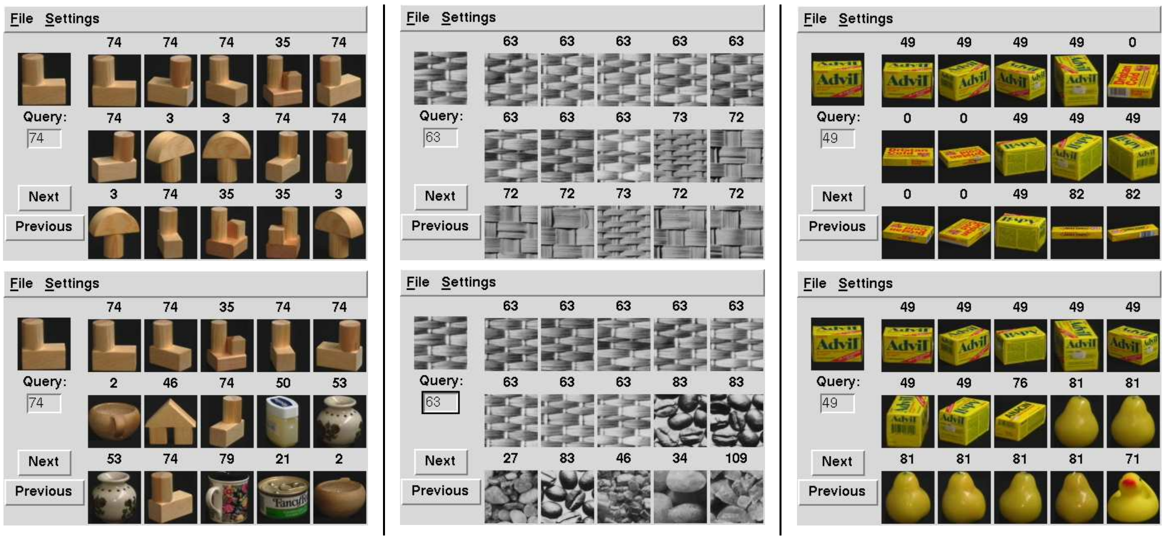


Figure 5. Comparison of EMM/ML retrieval results (top row) with those of HI on Columbia and MRSAR/MD on Brodatz (bottom row).

6.2 Invariance

As discussed in section 5, EMMs provide two ways to encode invariance into the retrieval operation: learning and low-pass filtering (discarding high frequency subspaces).

Figures 4 c) and d) present 1) the surface spanned by precision/recall as a function of the number of subspaces considered during retrieval, and 2) the impact on precision/recall of the spacing between adjacent features vectors taken into account during learning. The shape of the precision/recall surface illustrates the trade-off between invariance and spatial support. When too few subspaces are considered, there is not enough support to capture the correlations of each texture class. Performance therefore increases as the number of subspaces grows, starting to degrade again when we get to the region of high frequencies. At this point, because the representation is very detailed, good recognition requires precise alignment between the query and the database features. Overall, there is a large set of subspaces for which a good trade-off between spatial support and invariance is achieved and precise selection of the number of subspaces is, in practice, not crucial for good performance.

Plot d) shows how retrieval accuracy can be improved by increasing the variability of the examples in the training set. When image blocks are non-overlapping the representation is invariant only to translations by multiples of the block size. As the inter-sample spacing decreases, the representation becomes invariant to smaller displacements and retrieval accuracy increases. This was expected, since invariance is a bigger problem when high frequencies are included. Similar results were obtained with Columbia.

References

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

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

[3] W. N. et al. The QBIC project: Querying images by content using color, texture, and shape. In *Storage and Retrieval for Image and Video Databases*, pages 173–181, SPIE, Feb. 1993, San Jose, California.

[4] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image Understanding Using Color Correlograms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 762–768, San Juan, Puerto Rico, 1997.

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

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

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

[8] S. Sclaroff. Deformable Prototypes for Encoding Shape Categories in Image Databases. *Pattern Recognition*, 30(4), April 1997.

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

[10] N. Vasconcelos and A. Lippman. Embedded Mixture Modeling for Efficient Probabilistic Content-Based Indexing and Retrieval. In *SPIE Multimedia Storage and Archiving Systems III*, Boston, 1998.

[11] N. Vasconcelos and A. Lippman. Probabilistic Retrieval: New Insights and Experimental Results. In *Proc. IEEE Workshop on Content-based Access to Image and Video Libraries*, CVPR99, Fort Collins, Colorado, 1999.