FEATURE REPRESENTATIONS FOR IMAGE RETRIEVAL: BEYOND THE COLOR HISTOGRAM

Nuno Vasconcelos and Andrew Lippman

MIT Media Lab, {nuno,lip}@media.mit.edu

ABSTRACT

We study solutions to the problem of feature representation in the context of content-based image retrieval (CBIR). Retrieval is formulated as a classification problem, where the goal is to minimize probability of retrieval error. Under this formulation, retrieval performance is directly related to the quality of density estimation which is, in turn, determined by properties of the feature representation. We show that most representations of interest for the retrieval problem are particular cases of the mixture model, and present detailed arguments for why this is the most appropriate representation for retrieval.

1. INTRODUCTION

An architecture for image retrieval is composed by three fundamental building blocks: a feature transformation, a feature representation and a similarity function. We have recently introduced a formulation of retrieval as a classification problem, where the goal is to minimize the probability of retrieval error, and shown that most of the current similarity functions are particular cases of this formulation [9, 8].

Under this formulation, retrieval performance is determined to a significant extent by the quality of density estimates which, in turn, are strongly impacted by the selection of feature representation. We show that most representations of interest for the retrieval problem, including parametric and non-parametric densities, vector quantization and histograms, are particular cases of the mixture model. We then argue that the mixture representation is the most appropriate for image retrieval, and present retrieval results on a database of generic imagery where it is shown to outperform color histograms, color correlograms, and the standard representations for texture-based retrieval.

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

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

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

We set 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 for the i^{th} class and $P(y=i)$ its prior probability. The smallest achievable probability of error is the *Bayes error*

$$L^* = 1 - E_{\mathbf{x}}[\max_i P(y=i|\mathbf{x})], \qquad (3)$$

and the difference between the actual error an this optimal bound is a function of the quality of density estimates [1]

$$\begin{aligned} &P(g(\mathbf{x}) \neq y) - L^* \leq && (4) \\ &\quad \sum_{i=1}^{M} \int |P(\mathbf{x}|y=i)P(y=i) - \hat{p}(\mathbf{x}|y=i)\hat{p}(y=i)| d\mathbf{x}, \end{aligned}$$

where $\hat{p}(\mathbf{x}|y=i)$ and $\hat{p}(y=i)$ are the estimates for $P(\mathbf{x}|y=i)$ and $P(y=i)$. It is therefore clear that good density estimation is a sufficient condition for accurate retrieval.

3. FEATURE REPRESENTATIONS

Among the different components of a retrieval system, the model (or feature representation) on which density estimates are based has the strongest impact on their accuracy. In this section we analyze the relationships between several feature representations of interest for image retrieval.

3.1. Mixture models

A mixture density has the form

$$P(\mathbf{x}) = \sum_{i=1}^{C} P(\mathbf{x}|\omega_i)P(\omega_i), \tag{5}$$

where C is a number of classes, $\{P(\mathbf{x}|\omega_i)\}_{i=1}^{C}$ a sequence of *class-conditional densities*, and $\{P(\omega_i)\}_{i=1}^{C}$ a sequence of *class probabilities*. Mixture densities model processes with hidden structure: one among the C classes is first selected according to the $\{P(\omega_i)\}$, and the observed data is then drawn according to the respective class-conditional density. Class-conditional densities can be any valid probability density functions, i.e. any set of non-negative functions integrating to one. In this paper we consider the subset of mixture models where class-conditional densities are a function of two parameters: scale and location.

3.2. Parametric densities

It is obvious from (5) that any parametric density can be seen as a particular case of a mixture model, by simply making $C = 1$. In particular, for a Gaussian of mean μ and covariance $\mathbf{\Sigma}$

$$P(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n|\Sigma|}e^{-\frac{1}{2}||\mathbf{x}-\mu||_{\mathbf{\Sigma}}} \tag{6}$$

where

$$||\mathbf{x}-\mu||_{\mathbf{\Sigma}} = (\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu). \tag{7}$$

3.3. Non-parametric densities

It is also clear from (5) that given a sample of observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, by making the number of image classes equal to the number of observations M, assuming each class to be equally likely, and class conditional densities to be replicas of the same kernel $\mathcal{K}_\Sigma(\mathbf{x})$ centered on the observations

$$P(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} \mathcal{K}_\Sigma(\mathbf{x}-\mathbf{x}_i) \tag{8}$$

we obtain what are usually called, *Parzen* or *kernel* density estimates [6]. These models are traditionally referred to as non-parametric densities, even though they usually require the specification of a scale (or *bandwidth*) parameter Σ. One popular choice for the kernel $\mathcal{K}_\Sigma(\mathbf{x})$ is the Gaussian distribution, in which case Σ is a covariance matrix.

3.4. Vector quantization

In order to relate mixtures with vector quantization, we start by noticing that associated with any mixture model there is a soft partition of the sample space. In particular, given an observation $\mathbf{x}$, it is possible to assign that observation to each of the data classes according to

$$\begin{aligned} P(\omega_i|\mathbf{x}) &= \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{k=1}^{C} P(\mathbf{x}|\omega_k)P(\omega_k)} && (9) \\ &= \frac{1}{1+\sum_{k\neq i}\frac{P(\mathbf{x}|\omega_k)P(\omega_k)}{P(\mathbf{x}|\omega_i)P(\omega_i)}}. && (10) \end{aligned}$$

If the class-conditional densities are Gaussian then

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

Making all the covariances tend to zero, $\mathbf{\Sigma}_i = \lim_{\epsilon\to 0}\epsilon\mathbf{I}, \forall i$

$$P(\mathbf{x}|\omega_i) = \delta(\mathbf{x}-\mu_i), \tag{11}$$

where $\delta(\mathbf{x})$ is the Dirac delta function [5]. Therefore,

$$P(\mathbf{x}) = \sum_{i=1}^{C}\delta(\mathbf{x}-\mu_i)P(\omega_i), \tag{12}$$

and

$$\begin{aligned} P(\omega_i|\mathbf{x}) &= \frac{1}{1+\sum_{k\neq i}\frac{P(\omega_k)}{P(\omega_i)}e^{\frac{1}{\epsilon}(||\mathbf{x}-\mu_i||^2-||\mathbf{x}-\mu_k||^2)}} \\ &= \begin{cases} a, & \text{if } ||\mathbf{x}-\mu_i|| \leq ||\mathbf{x}-\mu_k|| \forall k < i \\ 0, & \text{otherwise.}, \end{cases} && (13) \end{aligned}$$

where

$$a = \frac{P(\omega_i)}{P(\omega_i) + \sum_{\{k:||\mathbf{x}-\mu_k|=||\mathbf{x}-\mu_i||\}}P(\omega_k)}.$$

Since the set $\{\mathbf{x} : ||\mathbf{x}-\mu_k|| = ||\mathbf{x}-\mu_i||\}$ has measure zero, for all practical purposes (13) can be approximated by

$$P(\omega_i|\mathbf{x}) = \begin{cases} 1, & \text{if } ||\mathbf{x}-\mu_i|| \leq ||\mathbf{x}-\mu_k|| \forall k < i \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

and some tie-breaking rule used to assign points that lie on the boundaries between different cells.

Equations (11) and (14) are a generative model for a vector quantizer [2]. The space is first partitioned into a collection of Voronoi cells according to (14), an operation usually referred to as *encoding* in the VQ literature. Each observed sample $\mathbf{x}$ is then assigned to the cell in which it falls and the cell's label transmitted to a *decoder*. Given a cell label, the decoder simply draws a sample from the conditional density associated with the cell, according to (11). Since this density is the delta function this simply consists of outputting the centroid or mean of the cell. It is, therefore, clear that a VQ is a particular case of the Gaussian mixture model.

3.5. Histograms

Equations (12) and (14) also provide an interpretation of a VQ as an histogram since, in practice, the vector $\mathbf{H} = \{P(\omega_1), \ldots P(\omega_C)\}$ is estimated with normalized counts of the number of samples that land on each of the quantization bins. Since any histogram can be expressed in this form, it is clear that histograms are also a particular case of the mixture model. A special case of interest occurs when the reproducing vectors μ_i are located on a rectangular grid of size $h_1, \ldots, h_n$. In this case, the Voronoi cells become rectangles

$$P(\omega_i|\mathbf{x}) = \begin{cases} 1, & \text{if } |\mathbf{x}_1 - \mu_{i,1}| \leq \frac{h_1}{2}, \ldots, |\mathbf{x}_n - \mu_{i,n}| \leq \frac{h_n}{2} \\ 0, & \text{otherwise,} \end{cases}$$

and we obtain the standard histogram model that is commonly used for color-based image retrieval [7].

4. A CRITICAL ANALYSIS

Since most of the current feature representations are particular cases of the mixture model, it is natural to expect that they will lead to suboptimal performance when applied to the retrieval problem. In this section, we give more concrete arguments for why this is indeed the case.

We start by noticing that because they have as many parameters as the number of observations (image features in the case of image retrieval), the so called non-parametric models are not a compact representation for the underlying density. Consequently, the evaluation of equation (8) is computationally expensive (complexity proportional to the number of training features). On the other hand there is no guarantee that the density estimates will be better than those provided by the mixture model, and there is usually no easy way to set the bandwidth parameter [6]. It is, therefore not clear that relying on a non-parametric model will justify the increase in retrieval complexity.

While non-parametric models have too many degrees of freedom, parametric ones have to few. Typically, retrieval is performed on databases of non-homogeneous images, composed by multiple visual stimulae, and the associated densities are only rarely unimodal. Consider the simple example of the image in Figure 1 a). Its intensity histogram, shown in b), has two main peaks: one for the black background and another for the white letters in the foreground. The fact that most parametric densities are unimodal makes it clear that, in general, they do not have enough expressive power to capture the details of the densities associated with real images such as this one. In particular, Figure 1 c) illustrates how bad a Gaussian fit can be.

While overcoming this lack of expressive power, standard color histograms also suffer from significant limitations. Because their complexity (number of cells) grows exponentially with the dimension of the feature space they are only practical in low dimensions, such as the 3D space of image colors. Consequently they have no ability to support the features with large spatial support that are required to model spatial image dependencies. Hence, they can only provide a very coarse image characterization that is insufficient for fine image discrimination. This is illustrated by Figure 1 d), where we present an image that has the exact same histogram as the one in a) but which is visually very different from it. Because this limitation is inherent to the fixed partition of the space associated with the standard histogram, most attempts to extend its spatial support, e.g. color coherence vectors and correlograms [3], are also plagued by the exponential dependence of complexity on the dimensionality of the space.

On the other hand, by adapting the partition of the space to the characteristics of the data, VQ achieves equivalent performance to the standard histogram with much lower complexity. This is not surprising since, as seen above, a VQ is a generic form of histogram. Still, the fact that VQ-based estimates rely on a hard partition of the space restricts their usefulness since slight feature perturbations may lead to drastic changes in quantization, label histograms and consequently image similarity. This effect is illustrated in Figure 2.

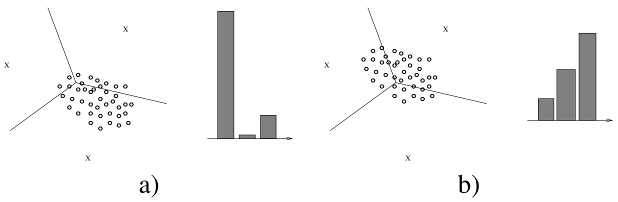
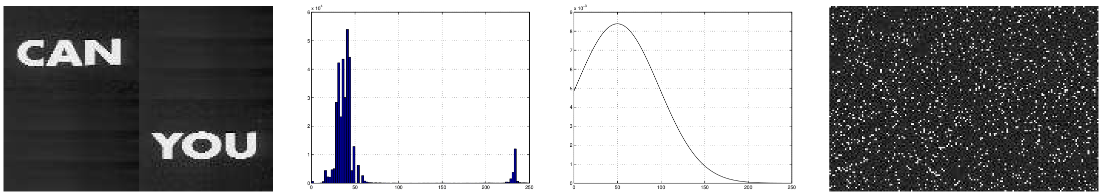


Fig. 2. a) Partition of the feature space by a 3-cell VQ, a set of feature vectors and the corresponding label histogram. b) Small perturbations of the feature vectors can lead to entirely different label histograms.

Because, as discussed above, mixture models perform a soft partition of the feature space, they are not subject to these problems. Furthermore, by allowing arbitrary covariances for each of the data classes, Gaussian mixtures provide a much better approximation to the true density than the train of delta functions associated with VQ. One can therefore conclude that there is no strong justification for relying on any of the above feature representations instead of the Gaussian mixture. In the following section we validate these theoretical arguments with experimental evidence.

5. EXPERIMENTAL RESULTS

In order to compare the performance achievable with different representations we conducted experiments on a database containing 15 image classes from Corel in a total of $1,500$



a) b) c) d)

Fig. 1. a) and d) two images with the color histogram shown in b). c) best Gaussian fit.

images. Four representations were tested: Gaussian, color histograms, Gaussian mixture, and color correlograms. The various representations are usually employed for different tasks. The Gaussian is implicit in the *Mahalanobis distance* (MD) traditionally used for texture-based retrieval while histograms are mostly used for color-based retrieval. Correlograms aim to capture both color and texture.

The Gaussian mixture model was applied to the space of coefficients of a block-wise 8×8 DCT transform. Since the low-dimensional projection of a Gaussian mixture is still a Gaussian mixture, the resulting density can be seen as an extension of the color histogram [8, 10]. When only the first coefficient is considered for retrieval, the two representations are identical. Including more coefficients extends the spatial support of the representation and should improve performance. The parameters of the two models were set so that they had equivalent retrieval complexity. Because the application of the MD to the DCT features lead to very poor performance we combined instead the Gaussian representation with a feature transformation that is more tuned for texture characterization: the MRSAR features [4]. This combination has been used in several retrieval systems. For correlograms we followed [3], once again picking the parameters that led to complexity equivalent to that of the Gaussian mixture. Figure 3 presents precision/recall curves for the various approaches, showing that the Gaussian mixture representation clearly outperforms the others.

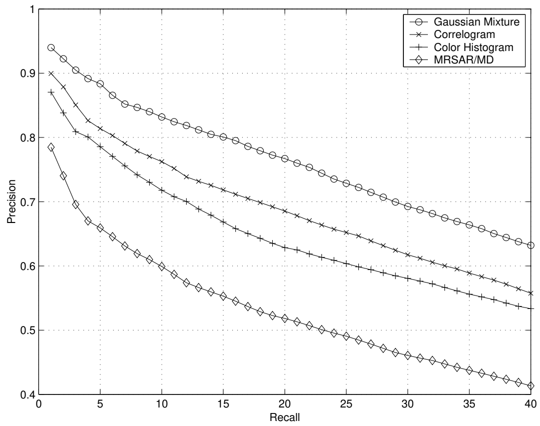


Fig. 3. Precision/recall on Corel.

6. REFERENCES

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

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

[3] 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.

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

[5] A. Papoulis. *The Fourier Integral and its Applications*. McGraw-Hill, 1962.

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

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

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

[9] N. Vasconcelos. A Unified View of Image Similarity. In *Proc. Int. Conf. Pattern Recognition*, Barcelona, Spain, 2000.

[10] 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.