A Unifying View of Image Similarity

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

Abstract

We study solutions to the problem of evaluating image similarity 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. It is shown that this formulation establishes a common ground for comparing similarity functions, exposes assumptions hidden behind most of the ones in common use, enables a critical analysis of their relative merits, and determines the retrieval scenarios for which each may be most suited. We conclude that most of the current similarity functions are sub-optimal special cases of the Bayesian criteria that results from explicit minimization of error probability.

1 Introduction

An architecture for image retrieval is composed by three fundamental building blocks: a feature transformation, a feature representation and a similarity function. In this paper, we study good similarity criteria. Given the long history of similarity evaluation in fields such as texture or object recognition, it is not surprising that many similarity functions have been proposed for the retrieval problem. However, there seems to be no clear understanding of their inter-relationships or relative weaknesses and strengths. In practice, similarity functions are frequently selected without any justification or consideration for the underlying assumptions. For example, while the Gaussian assumption underlying quadratic metrics such as the *Mahalanobis distance* (MD) is acceptable for the homogeneous images that compose most texture databases, it is inappropriate when dealing with generic imagery. Yet, the Mahalanobis distance is frequently used for generic image retrieval [4, 7].

In this paper, we formulate retrieval as a classification problem, where the goal is to minimize the probability of retrieval error. This is a generic criteria, sensible independently of the particular type of images that populate a database. By making explicit the assumptions behind several similarity functions in current use, the new formulation enables a critical analysis of their relative merits. We point out that the explicit minimization of probability of error leads to a Bayesian retrieval criteria, and show that most of the current similarity functions are sub-optimal special cases of it. These theoretical findings are confirmed experimentally for texture and color-based retrieval, where the Bayesian criteria is shown to outperform the most popular solutions for these tasks: MD and *histogram intersection* (HI) [10].

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

$$g : \mathcal{F} \to M = \{1, \ldots, K\}$$

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 [5]

$$g^*(\mathbf{x}) = \arg\max_i P(\mathbf{x}|y = i)P(y = i), \qquad (1)$$

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* [5]

$$L^* = 1 - E_{\mathbf{x}}[\max_i P(y = i|\mathbf{x})]. \qquad (2)$$

We next demonstrate that most of the similarity functions in current use for image retrieval are special cases of the Bayesian criteria.

3 Unifying similarity evaluation

The relations between various similarity functions are illustrated in Figure 1. If an upper bound on the Bayes error of a collection of two-way classification problems is minimized instead of the probability of error of the original problem, the Bayesian criteria reduces to the *Bhattacharyya distance* (BD). On the other hand, if the original criteria is minimized, but the different image classes are assumed to be equally likely a priori, we have the *maximum likelihood* (ML) retrieval criteria. As the number of query points grows to infinity the ML criteria tends to the *Kullback-Leibler divergence* (KLD), which in turn can be approximated by the χ^2 test by performing a simple 1^{st} order Taylor series expansion. Alternatively, the KLD can be simplified by assuming that the underlying probability densities belong to a pre-defined family. In the Gaussian case, further assumption of orthonormal covariance matrices leads to the *quadratic distance* (QD) frequently found in the compression literature. The next possible simplification is to assume that all classes share the same covariance matrix, leading to the *Mahalanobis distance* (MD). Finally, assuming identity covariances results in the *Euclidean distance* (ED). We next derive in greater detail all these relationships.

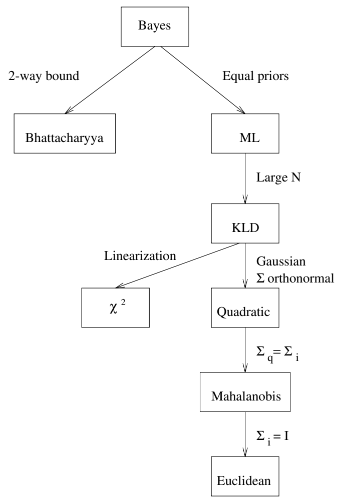


Figure 1. Relations between different similarity functions.

3.1 Bhattacharyya distance

If there are only two classes in the classification problem, (2) can be written as [5]

$$\begin{aligned} L^* &= E_{\mathbf{x}}[\min(P(y=0|\mathbf{x}), P(y=1|\mathbf{x})] \\ &= \int \min[P(\mathbf{x}|y=0)P(y=0), P(\mathbf{x}|y=1)P(y=1)]dx \\ &\leq \sqrt{P(y=0)P(y=1)} \int \sqrt{P(\mathbf{x}|y=0)P(\mathbf{x}|y=1)}dx \end{aligned}$$

where we have used the bound $\min[a, b] \leq \sqrt{ab}$. The last integral is usually known as the Bhattacharyya distance between $P(\mathbf{x}|y=0)$ and $P(\mathbf{x}|y=1)$ and has been proposed (e.g. [8, 2]) for image retrieval, where it takes the form

$$g(\mathbf{x}) = \arg\min_i \int \sqrt{P(\mathbf{x}|q)P(\mathbf{x}|y=i)}, \qquad (3)$$

and $P(\mathbf{x}|q)$ is the density of the query. The resulting classifier can thus be seen as the one which finds the lowest upper-bound on the Bayes error for the collection of two-class problems involving the query and each of the database classes.

3.2 Maximum likelihood

It is straightforward to see that when all image classes are a priori equally likely, $P(y=i) = 1/K$, (1) reduces to the ML classifier

$$g(\mathbf{x}) = \arg\max_i P(\mathbf{x}|y=i). \qquad (4)$$

Under the assumption that the query consists of a collection of N independent query features $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, this equation can also be written as

$$g(\mathbf{x}) = \arg\max_i \frac{1}{N} \sum_{j=1}^{N} \log P(\mathbf{x}_j|y=i). \qquad (5)$$

3.3 Kullback-Leibler divergence

When the number of query features N is large, simple application of the law of large numbers to (5) reveals that

$$\begin{aligned} g(\mathbf{x}) &\stackrel{N\to\infty}{\longrightarrow} \arg\max_i E_q[\log P(\mathbf{x}|y=i)] \\ &= \arg\max_i \int P(\mathbf{x}|q) \log P(\mathbf{x}|y=i)dx \\ &= \arg\min_i \int P(\mathbf{x}|q) \log \frac{P(\mathbf{x}|q)}{P(\mathbf{x}|y=i)}dx \\ &= \arg\min_i KL(Q||P_i) \end{aligned}$$

where $KL(Q||P_i)$ the Kullback-Leibler divergence between the query density and that associated with the i^{th} database image class [3]. Thus, the KLD is simply the asymptotic limit of the ML criteria.

3.4 χ^2 statistic

Using the first order Taylor series approximation for the log about $x = 1$, $\log(x) \approx x - 1$, we obtain

$$KL(q||p) \approx \int \frac{q(x)^2 - q(x)p(x)}{p(x)}dx$$

$$= \int \left(\frac{q(x)^2 - q(x)p(x)}{p(x)} - q(x) + p(x) \right) dx$$
$$= \int \frac{(q(x) - p(x))^2}{p(x)} dx$$

and, in particular,

$$KL(Q||P_i) \approx \int \frac{(P(\mathbf{x}|q) - P(\mathbf{x}|y=i))^2}{P(\mathbf{x}|y=i)} dx \quad (6)$$

The integral on the right is known as the χ^2 statistic and has been proposed as a metric for image similarity in [9, 1]. It is clear that it simply approximates the KLD and, therefore, the ML classifier in the asymptotic limit of a large query.

3.5 Quadratic distance

When the image features are Gaussian distributed

$$P(\mathbf{x}|y=i) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}_i|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)} \quad (7)$$

(5) becomes

$$g(\mathbf{x}) = \arg\min_i \log|\mathbf{\Sigma}_i| + \hat{\mathcal{L}}_i \quad (8)$$

where $\hat{\mathcal{L}}_i = \frac{1}{N}\sum_n (\mathbf{x}_n - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mu_i)$ is the quadratic distance commonly found in the compression literature. Thus, as a retrieval metric, the QD can be seen as the result of imposing two stringent restrictions on the ML criteria. First, that all image sources are Gaussian and, second, that their covariance matrices are orthonormal ($|\mathbf{\Sigma}_i| = 1, \forall i$).

3.6 Mahalanobis distance

Because

$$\begin{aligned}
\hat{\mathcal{L}}_i = \frac{1}{N}\sum_n (\mathbf{x}_n - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mu_i) = \\
&= \frac{1}{N}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}} + \hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \hat{\mathbf{x}} + \hat{\mathbf{x}} - \mu_i) \\
&= \frac{1}{N}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}})^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \hat{\mathbf{x}}) \\
&\quad -2(\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1} \frac{1}{N}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}}) + \\
&\quad (\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\hat{\mathbf{x}} - \mu_i) \\
&= \frac{1}{N} trace[\mathbf{\Sigma}_i^{-1} \sum_n (\mathbf{x}_n - \hat{\mathbf{x}})(\mathbf{x}_n - \hat{\mathbf{x}})^T] + \mathcal{M}_i \\
&= trace[\mathbf{\Sigma}_i^{-1} \hat{\mathbf{\Sigma}}_{\mathbf{x}}] + \mathcal{M}_i \quad (9)
\end{aligned}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{x}}$ is the sample covariance of $\mathbf{x}_n$ and $\mathcal{M}_i = (\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\hat{\mathbf{x}} - \mu_i)$ the MD, this metric results from complementing Gaussianity with the assumption that all classes have the same covariance ($\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{\Sigma}_i = \mathbf{\Sigma}, \forall i$). Finally, if this covariance is the identity ($\mathbf{\Sigma} = \mathbf{I}$), we obtain the square of the Euclidean Distance $\mathcal{E}_i = (\hat{\mathbf{x}} - \mu_i)^T(\hat{\mathbf{x}} - \mu_i)$.

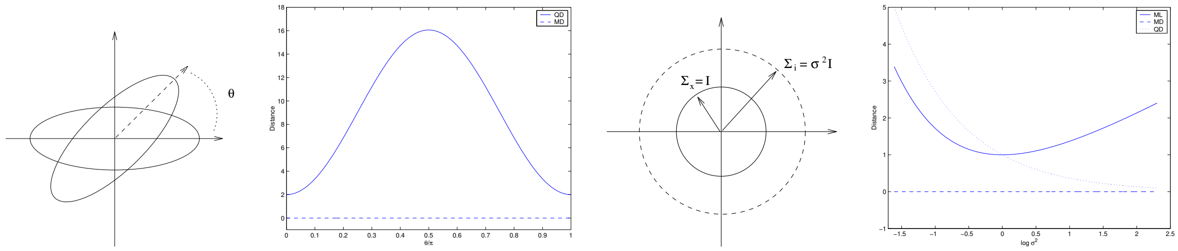
4 A critical analysis

Exposing the assumptions behind each similarity function enables a critical analysis of their usefulness and the determination of the retrieval scenarios for which they may be most appropriate. While the choice between the Bayesian and ML criterion is a function only of the amount of prior knowledge about class probabilities, there is in general no strong justification to rely on any of the remaining metrics.

For example, while ML and KLD perform equally well for *global queries* based on entire images, ML has the added advantage of also enabling *local queries* consisting of user-selected image regions. These queries are important because they allow users to express exactly what interests them within a retrieval image and, therefore, are considerably less ambiguous than global queries. The only advantage of the KLD is a smaller computational complexity, whenever global queries are enough and it has a closed-form expression. This is also true for the χ^2 statistic which, unlike the KLD, is not guaranteed to equal the performance of ML even for large queries. Finally, relying on the Bhattacharyya distance seems even less sensible. There is small justification to replace the minimization of the error probability on the multi-class retrieval problem (ML) by the search for the two class problem with the smallest error bound (BD).

The remaining retrieval criteria (QD, MD, and ED) only make sense if the image features are Gaussian distributed for all classes. While this is approximately true in certain cases (e.g. texture databases where each image is an homogeneous patch of a given texture) it certainly does not hold for generic databases. Even when the Gaussian approximation holds, there seems to be little justification to rely on any criteria other than ML, as all other are approximations that arbitrarily discard covariance information. This information is usually important for the detection of subtle variations such as rotation and scaling in feature space. This is illustrated by Figure 2. In a) and b) we show the distance, under both QD and MD between a Gaussian and a replica rotated by $\theta \in [0, \pi]$. Plot b) clearly illustrates that while the MD has no ability to distinguish between the rotated Gaussians, the inclusion of the $trace[\mathbf{\Sigma}_i^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}}]$ term leads to a much more intuitive measure of similarity: minimum when both Gaussians are aligned and maximum when they are rotated by 90^o.

As illustrated by c) and d), further inclusion of the term $\log|\mathbf{\Sigma}_i|$ (full ML retrieval) penalizes mismatches in scaling. In plot c) we show two Gaussians, with covariances $\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{I}$ and $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$, centered on zero. In this ex-



a) b) c) d)

Figure 2. a) a Gaussian with mean $(0, 0)^T$ and covariance $diag(4, 0.25)$ and its replica rotated by θ. b): Distance between the Gaussian and its rotated replicas as a function of θ/π under both the QD and the MD. c) two Gaussians with different scales. d) Distance between them as a function of $\log \sigma^2$ under ML, QD, and MD.

ample, MD is always zero, while $trace[\mathbf{\Sigma}_i^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}}] \propto 1/\sigma^2$ penalizes small σ and $\log |\mathbf{\Sigma}_i| \propto \log \sigma^2$ penalizes large σ. The total distance is shown as a function of $\log \sigma^2$ in plot d) where, once again, we observe an intuitive behavior: the penalty is minimal when both Gaussians have the same scale ($\log \sigma^2 = 0$), increasing monotonically with the amount of scale mismatch. Notice that if the $\log |\mathbf{\Sigma}_i|$ term is not included, large changes in scale may not be penalized at all.

5 Experimental results

Figure 3, presents precision/recall curves for texture and color-based retrieval experiments on (respectively) the Brodatz (texture) and Columbia (object) databases. Two curves are presented for each database, one relative to ML and another relative to the similarity function commonly used for the associated task: MD for texture and HI for color. On Brodatz, texture features are the coefficients of the least squares fit to the MRSAR model [6]. On Columbia, we use color histograms of 512 bins on YUV space. To make comparisons fair, on Brodatz ML is based on the Gaussian assumption of (8).

The figure confirms that there is a clear improvement in switching from MD or HI to ML retrieval. For Brodatz the gain is approximately uniform and always between $5-10\%$. On Columbia, both metrics perform equally well for low recall, but ML has substantially higher precision (up to 15% better) for high recall.

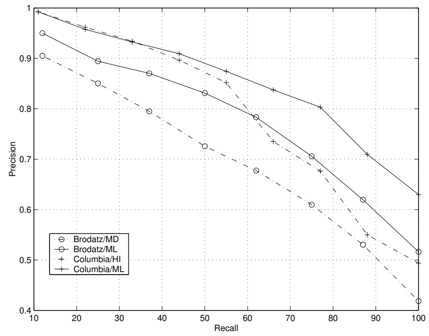


Figure 3. Precision/recall for texture and color retrieval.

References

[1] J. D. Bonet, P. Viola, and J. F. III. Flexible Histograms: A Multiresolution Target Discrimination Model. In E. G. Zelnio, editor, *Proceedings of SPIE*, volume 3370, 1998.

[2] D. Comaniciu, P. Meer, K. Xu, and D. Tyler. Retrieval Performance Improvement through Low Rank Corrections. In *Workshop in Content-based Access to Image and Video Libraries*, pages 50–54, 1999, Fort Collins, Colorado.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[4] 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.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

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

[7] T. Minka and R. Picard. Interactive learning using a "society of models". *Pattern Recognition*, 30:565–582, 1997.

[8] B. Moghaddam, H. Bierman, and D. Margaritis. Defining Image Content with Multiple Regions-of-Interest. In *Workshop in Content-based Access to Image and Video Libraries*, pages 89–93, 1999, Fort Collins, Colorado.

[9] B. Schiele and J. Crowley. Object Recognition Using Multidimensional Receptive Field Histograms. In *Proc. 4^{th} European Conference on Computer Vision*, 1996, , Cambridge, UK.

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