Chapter 2

Retrieval as statistical inference

The central component of an architecture for content-based image retrieval (CBIR) is a criteria for evaluating image similarity. Typically, this is achieved by defining a similarity function that maps the space of image classes that compose a database into the space of possible orderings for those classes. In this chapter, we argue that a natural goal for a retrieval system is to minimize the probability of retrieval error[1]. This leads to a new formulation of the retrieval problem, derived from Bayesian decision theory, and a probabilistic criteria for the evaluation of image similarity.

In addition to minimizing retrieval error the Bayesian solution unifies a large body of similarity functions in current use. In particular, it is shown that most of these functions can be derived from Bayesian retrieval by 1) making assumptions with respect to the densities of the image classes or 2) approximating the quantities involved in Bayesian inference. This suggests that, even if minimizing probability of error is not the desired goal for the retrieval system, there is no apparent reason to prefer those functions to the Bayesian counterpart. The theoretical claims are validated through retrieval experiments that confirm the superiority of Bayesian similarity.

[1] A more generic performance criteria is the Bayes risk [10] where different types of errors are assigned different costs. Because we currently do not have good strategies to define such costs, we simply assign a unitary cost to all errors (and zero cost to all correct decisions), in which case Bayes risk is equivalent to the probability of error. It would, however, be straightforward to extend the retrieval formulation presented in the thesis to the minimization of Bayes risk, if more detailed costs were available.

2.1 Terms and notation

We start by defining some terms and notation. An *image* I is a map from a two-dimensional pixel lattice of size $P \times Q$

$$\mathcal{L} = \{1, \ldots, P\} \times \{1, \ldots, Q\} \tag{2.1}$$

into the space $\mathcal{A}$ of all $P \times Q$ arrays of pixel colors

$$I : \mathcal{L} \rightarrow \mathcal{A}.$$

The *color* of pixel $(i, j) \in \mathcal{L}$ is denoted by $I_{i,j}$ and can be a scalar (for gray-scale images) or a 3-D vector (for color images). In the former case, the pixel color is also referred to as *intensity*. The number of color channels in an image is denoted by c.

We define two indicator functions. For any set E, the *set indicator* function is

$$\chi_E(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

For any two integers i and j, the *Kronecker delta* function is defined by

$$\delta_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

A *partition* of a set E is a collection of subsets (also known as *partition cells* or *regions*) $\{E_1, \ldots, E_R\}$ that are disjoint and cover E, i.e.

$$\cup_{i=1}^{R} E_i = E \text{ and } E_i \cap E_j = \emptyset, \ \forall i \neq j. \tag{2.4}$$

An *image database* $\mathbf{D}$ is a collection of images

$$\mathbf{D} = \{I_1, \ldots, I_S\}$$

where S is the *database size*. Within a database, images are organized into M *image classes*

$$\mathbf{D} = \{\mathbf{D}_1, \ldots, \mathbf{D}_M\}$$

where the $\mathbf{D}_i$ are a partition for $\mathbf{D}$.

In general, a classification of the images in the database is available. If that is not the case, two alternatives can be pursued. The first is to assume that each image defines a class by its own. This solution reflects the absence of any prior knowledge about the database content and leads to as many classes as the cardinality of the database. We denote this type of structure as a *flat database*. The second is to try to generate the classification either automatically or manually. Since individual images can always be seen as subclasses inside the classes $\mathbf{D}_i$ we call this organization a *hierarchical database*. Of course, there can be multiple levels in the hierarchical organization of a database. *In all the theoretical derivations of the thesis we assume that the images are already classified. For experiments we always rely on a flat database structure.* The issue of automatically grouping the images in the database, or *indexing*, is not addressed.

Associated with an image database there is a space $\mathcal{Z} \subset R^n$ of *image observations*. An image observation $\mathbf{z} = \{z_1, \ldots z_n\}$ is a vector containing n pixel colors extracted from an image. The *region of support* of observation $\mathbf{z}$ is the set of pixels in $\mathcal{L}$ whose colors are represented in $\mathbf{z}$. It can be a single pixel $(n = c)$ or any number b of them $(n = cb)$. When $b > 1$, the regions of support of different observations can overlap and, consequently, there can be as many observations as there are pixels in the image. A *feature transformation* is a map

$$T : \mathcal{Z} \rightarrow \mathcal{X}$$

from the space of image observations into some other space $\mathcal{X}$ deemed more appropriate to the retrieval operation. We call $\mathcal{X}$ the *feature space*, and $\mathbf{x} = T(\mathbf{z})$ a *feature vector*. *Features* are the elements of a feature vector and feature vectors inherit the region of support of the observations from which they are derived. If the feature transformation is the identity, then $\mathcal{Z}$ and $\mathcal{X}$ are the same.

A *feature representation* is a probabilistic model for how each of the image classes populates the feature space $\mathcal{X}$. We introduce a *class indicator* variable $Y \in \{1, \ldots, M\}$ and denote the *class-conditional probability density function (pdf)* or *class-conditional likelihood* associated with class i by $P_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = i)$. This can be any non-negative function integrating to one. Throughout the thesis, we use upper case for random variables and lower case for particular values, e.g. $\mathbf{X} = \mathbf{x}$ denotes that the random variable $\mathbf{X}$ takes the value $\mathbf{x}$. When the meaning is clear from context, we usually omit one of the symbols. For example, $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ is commonly used instead of $P_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = i)$. Boldface type is used to represent vectors.

One density that we will encounter frequently is the Gaussian, defined by a mean vector μ and a positive-definite covariance matrix $\mathbf{\Sigma}$ according to

$$\mathcal{G}(\mathbf{x}, \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} e^{-\frac{1}{2}||\mathbf{x}-\mu||^2_{\mathbf{\Sigma}}} \tag{2.5}$$

where

$$||\mathbf{x} - \mu||_{\mathbf{\Sigma}} = (\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \tag{2.6}$$

is the quadratic norm defined by $\mathbf{\Sigma}^{-1}$. The Euclidean norm is the particular case in which $\mathbf{\Sigma} = \mathbf{I}$. When $\mathbf{\Sigma} = \sigma \mathbf{I}$ and $\sigma \to 0$ the Gaussian converges to the Dirac function [123] defined by

$$\int \delta(\mathbf{x} - \mathbf{x}_0) f(\mathbf{x}) d\mathbf{x} = f(\mathbf{x}_0), \tag{2.7}$$

for all continuous functions $f(\mathbf{x})$.

Together, a *feature transformation* and a *feature representation* determine an *image representation*. An *image representation* and a *similarity function* define a *retrieval system*. This is a system that accepts queries from a user and searches a database for images that best match those queries. A *visual query* $\mathbf{x}$ is a collection of N feature vectors $\{\mathbf{x}_j\}_{j=1}^N$ extracted from a *query image*. If the the union of the regions of support of these feature vectors covers the entire lattice $\mathcal{L}$ the query is denoted as *global*. Otherwise, it is denoted as *local*. Local queries can be assembled through a graphical interface, by allowing a user to select a region or collection of regions from the query image. Throughout the thesis we rely on the following independence assumptions.

Assumption 1 *The feature vectors $\{\mathbf{x}_j\}_{j=1}^N$ included in a visual query are independent and identically distributed (iid)*

$$P_{\mathbf{X}_1,\ldots,\mathbf{X}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{j=1}^N P_{\mathbf{X}}(\mathbf{x}_j).$$

Assumption 2 *Given the knowledge of the true image class the query feature vectors $\{\mathbf{x}_j\}_{j=1}^N$ are independent*

$$P_{\mathbf{X}_j|Y,\mathbf{X}_1\ldots\mathbf{X}_{j-1},\mathbf{X}_{j+1},\ldots,\mathbf{X}_N}(\mathbf{x}_j|i, \mathbf{x}_1 \ldots \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_N) = P_{\mathbf{X}|Y}(\mathbf{x}_j|i)$$

By application of the chain rule of probability, Assumption 2 is equivalent to

$$P_{\mathbf{X}_1 \ldots \mathbf{X}_N|Y}(\mathbf{x}_1 \ldots \mathbf{x}_N|i) = \prod_{j=1}^{N} P_{\mathbf{X}|Y}(\mathbf{x}_j|i) \tag{2.8}$$

Given these definitions, we are now ready to address the questions posed by the design of a retrieval system. We start by considering the question of image similarity.

2.2 A Bayesian criteria for image similarity

In the image retrieval context, image similarity is naturally formulated as a problem of statistical classification. Given the feature space $\mathcal{X}$, a retrieval system is simply a map

$$\begin{array}{rccc} g: & \mathcal{X} & \to & \{1,\ldots,M\} \\ & \mathbf{x} & \mapsto & y \end{array}$$

from $\mathcal{X}$ to the index set of the M classes in the database. It is relatively common, in the vision and retrieval literatures, to define this map up-front without a clear underlying justification. For example, the most popular retrieval solution is to minimize the distance between color histograms[2] [172, 72, 49, 96, 2, 121, 193, 168, 149, 126, 43, 167, 163]. It is not clear that when confronted with the question "what would you like a retrieval system to do?" a naive user would reply "minimize histogram distance." In this work we define a more intuitive goal, the *minimization of probability of retrieval error*; i.e. we design systems that strive to be wrong as rarely as possible.

Definition 1 *A retrieval system is a map*

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

that minimizes

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

[2] We will give a precise definition of the term color histogram later on.

i.e. the system that has the minimum probability of returning images from a class $g(\mathbf{x})$ *different than that to which the query* $\mathbf{x}$ *belongs.*

Formulating the problem in this way has various advantages. First, the desired goal is stated explicitly, making clear what the retrieval operation is trying to achieve. Second, the criteria is objective leading to concrete metrics for evaluating the retrieval performance. Finally, it allows us to build on a relatively good theoretical understanding of the properties of various types of solutions (e.g. if their performance converges or not to that of the optimal solution and how quickly it does so) that are already in place for similar problems. In fact, once the problem is formulated in this way, the optimal solution is well known [38, 39, 48].

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*

$$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}). \tag{2.9}$$

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

$$L^* = 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \tag{2.10}$$

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

Proof: The proof can be found in various textbooks (see [38, 48] among many others). We include it here because 1) it is simple, and 2) provides insights for some later results.

The probability of error associated with the decision rule $g(\mathbf{x})$ is

$$P_{\mathbf{X},Y}(g(\mathbf{X}) \neq Y) = \int P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x})P_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = E_{\mathbf{x}}[P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x})], \tag{2.11}$$

where

$$\begin{aligned} P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) &= \sum_i P(Y \neq g(\mathbf{X})|\mathbf{X} = \mathbf{x}, Y = i)P_{Y|\mathbf{X}}(i|\mathbf{x}) \\ &= \sum_i (1 - \delta_{g(\mathbf{x}),i})P_{Y|\mathbf{X}}(i|\mathbf{x}) \\ &= 1 - \sum_i \delta_{g(\mathbf{x}),i}P_{Y|\mathbf{X}}(i|\mathbf{x}) \end{aligned} \tag{2.12}$$

and $\delta_{i,j}$ is the Kronecker delta function defined in (2.3). It follows that

$$\begin{aligned} P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x}) &\geq 1 - \max_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \\ &= 1 - P_{Y|\mathbf{X}}(Y = g^*(\mathbf{x})|\mathbf{x}) \\ &= P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x}) \end{aligned}$$

and, consequently

$$E_{\mathbf{x}}[P_{Y|\mathbf{X}}(Y \neq g(\mathbf{x})|\mathbf{x})] \geq E_{\mathbf{x}}[P_{Y|\mathbf{X}}(Y \neq g^*(\mathbf{x})|\mathbf{x})].$$

I.e., any other decision rule will have a larger probability of error than the Bayes classifier. Since, from (2.11),

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

the probability of error can never be smaller than the Bayes error. □

The *posterior probabilities* $P_{Y|\mathbf{X}}(i|\mathbf{x})$ are in general not easy to compute, making the direct implementation of this theorem difficult. To cope with this difficulty, several alternative approaches to the classification problem have been proposed in the now extensive classification literature. At the coarsest level, one can divide them into two major categories: *discriminant classifiers* and *classifiers based on generative models*.

Discriminant classifiers strive to find the surfaces in $\mathcal{X}$ that better separate the regions associated with the different classes in the sense of Theorem 1, classifying each point according to its position relative to those surfaces. Examples in this set are linear discriminant classifiers [39], neural networks [13], decision trees [18], and support vector machines [184], among others. From the retrieval point of view, discriminant classifiers have very limited interest because they must be retrained every time an image class is added to or deleted from the database. This is a strong restriction in the retrieval scenario, where databases can change daily or at an even faster pace.

Instead of dealing directly with (2.9), classifiers based on generative models take the alternative route provided by Bayes rule,

$$P_{Y|\mathbf{X}}(i|\mathbf{x}) = \frac{P_{\mathbf{X}|Y}(\mathbf{x}|i)P_Y(i)}{P_{\mathbf{X}}(\mathbf{x})}, \tag{2.13}$$

which leads to

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

When the query feature vectors $\{\mathbf{x}_j\}$ are iid, from (2.8)

$$\begin{aligned} g^*(\mathbf{x}) &= \arg\max_i \prod_{j=1}^{N} P_{\mathbf{X}|Y}(\mathbf{x}_j|i)P_Y(y=i) \\ &= \arg\max_i \sum_{j=1}^{N} \log P_{\mathbf{X}|Y}(\mathbf{x}_j|i) + \log P_Y(i), \end{aligned} \quad (2.14)$$

where $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ is the class-conditional likelihood for the i^{th} class and $P_Y(i)$ a *prior probability* for this class.

In the recent past, this similarity function has become prevalent for the evaluation of speech similarity and achieved significant success in tasks such as speech recognition and speaker identification [140, 145]. This is interesting because, if we can show that it also has good properties for visual similarity, we will have a common framework for dealing with images and sound. Also, because the individual likelihood functions $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ are learned for each image class independently, these classifiers can adapt easily to class additions and deletions. We denote (2.14) by *Bayesian retrieval criteria* and will refer to image retrieval based on it as *Bayesian retrieval*, *probabilistic retrieval*, or retrieval based on *Bayesian similarity*.

In practice, the probability of error of Bayesian retrieval is usually larger than the Bayes error. This is due to the fact that we do not know the true likelihood function or prior for each classes, and these have to be estimated from 1) images available in the database and 2) prior knowledge about the retrieval problem. We will return to this point in Chapter 3. For now, we analyze the relationships between Bayesian similarity and the similarity functions that are commonly used for image retrieval.

2.3 A unified view of image similarity

Figure 2.1 illustrates how various similarity functions commonly used for image retrieval are special cases of the Bayesian retrieval. While these functions do not exhaust the set of decisions rules that can be derived from or shown to be sub-optimal when compared to the Bayesian criteria (see chapter 3 of [38] for several others), we concentrate on them for two reasons: 1) they *have been* proposed as similarity functions, and 2) when available, derivations of their relationships to Bayesian similarity are scattered around the literature.

The figure illustrates that, 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 vectors grows to infinity the ML criteria tends to the *minimum discrimination information* (MDI), which in turn can be approximated by the χ^2 test by performing a simple first order Taylor series expansion. Alternatively, MDI can be simplified by assuming that the underlying probability densities belong to a pre-defined family. For *auto-regressive sources* it reduces to the *Itakura-Saito* distance that has received significant attention in the speech literature. 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 square of the *Euclidean distance* (ED). We next derive in more detail all these relationships.

2.3.1 Bhattacharyya distance

If there are only two classes in the classification problem, (2.10) can be written as [48]

$$\begin{aligned}
L^* &= E_{\mathbf{x}}[\min(P_{Y|\mathbf{X}}(0|\mathbf{x}), P_{Y|\mathbf{X}}(1|\mathbf{x}))] \\
&= \int P_{\mathbf{X}}(\mathbf{x}) \min[P_{Y|\mathbf{X}}(0|\mathbf{x}), P_{Y|\mathbf{X}}(1|\mathbf{x})] d\mathbf{x} \\
&= \int \min[P_{\mathbf{X}|Y}(\mathbf{x}|0)P_Y(0), P_{\mathbf{X}|Y}(\mathbf{x}|1)P_Y(1)] d\mathbf{x} \\
&\leq \sqrt{P_Y(0)P_Y(1)} \int \sqrt{P_{\mathbf{X}|Y}(\mathbf{x}|0)P_{\mathbf{X}|Y}(\mathbf{x}|1)} d\mathbf{x},
\end{aligned}$$

where we have used the bound $\min[a,b] \leq \sqrt{ab}$. The last integral is usually known as the Bhattacharyya distance between $P_{\mathbf{X}|Y}(\mathbf{x}|0)$ and $P_{\mathbf{X}|Y}(\mathbf{x}|1)$ and has been proposed (e.g. [111,

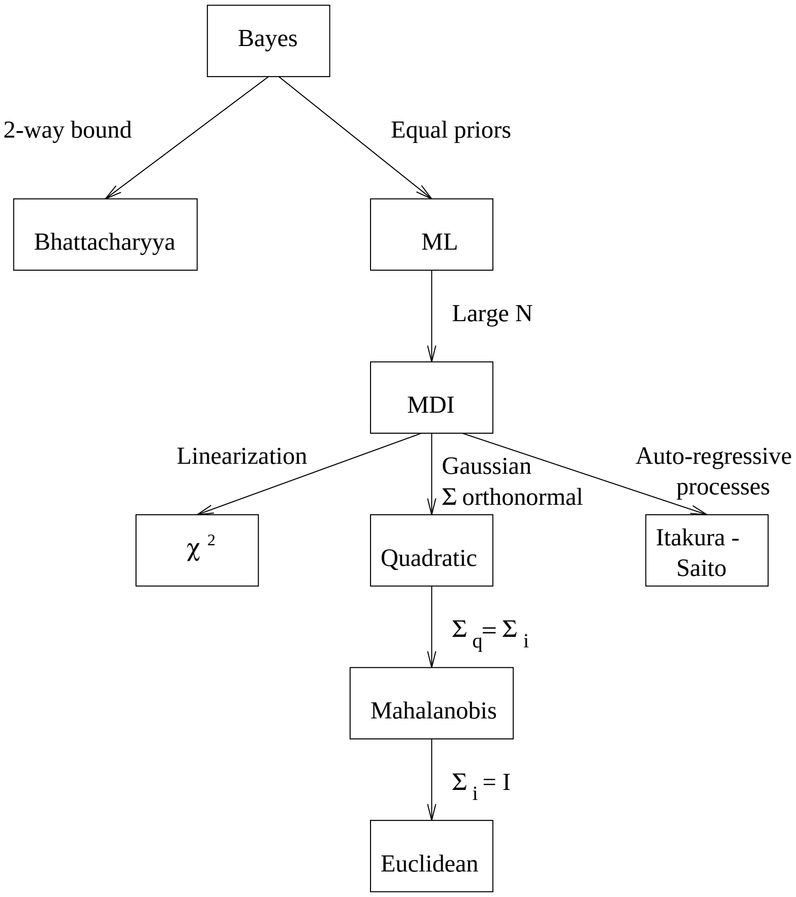

Figure 2.1: Relations between different image similarity functions.

30]) for image retrieval where, for a query density $P_{\mathbf{X}}(\mathbf{x})$, it takes the form

$$g(\mathbf{x}) = \arg\min_i \int \sqrt{P_{\mathbf{X}}(\mathbf{x}) P_{\mathbf{X}|Y}(\mathbf{x}|i)} d\mathbf{x}. \tag{2.15}$$

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.

Whenever it is possible to solve the minimization of the error probability on the multi-class retrieval problem it makes small sense to replace it by the search for the two class problem with the smallest error bound. Consequently, the above interpretation of the BD makes it clear that, in general, there is small justification to prefer it to Bayesian retrieval.

2.3.2 Maximum likelihood

It is straightforward to see that when all image classes are equally likely a priori, $P_Y(i) = 1/M$, (2.14) reduces to

$$g(\mathbf{x}) = \arg\max_i \frac{1}{N} \sum_{j=1}^{N} \log P_{\mathbf{X}|Y}(\mathbf{x}_j|i). \tag{2.16}$$

This decision rule is usually referred to as the maximum likelihood classifier. While, as we will see after Chapter 6, class priors $P_Y(i)$ provide a useful mechanism to 1) account for the context in which the retrieval operation takes place, 2) integrate information from multiple content modalities that may be available in the database, and 3) design learning algorithms, in Chapters 2-6 we assume that there is no a priori reason to prefer any given image over the rest. In this case, Bayesian and maximum likelihood retrieval are equivalent and we will use the two terms indiscriminately.

2.3.3 Minimum discrimination information

If $H_i, i = 1, 2$, are the hypotheses that $\mathbf{x}$ is drawn from the statistical population with density $P_i(\mathbf{x})$, the *Kullback-Leibler divergence* (KLD) or *relative entropy* [83, 31]

$$KL[P_2(\mathbf{x})||P_1(\mathbf{x})] = \int P_2(\mathbf{x}) \log \frac{P_2(\mathbf{x})}{P_1(\mathbf{x})} d\mathbf{x} \tag{2.17}$$

measures the mean information per observation from $P_2(\mathbf{x})$ for discrimination in favor of H_2 against H_1. Because it measures the difficulty of discriminating between the two populations, and is always non-negative and equal to zero only when $P_1(\mathbf{x}) = P_2(\mathbf{x})$ [83], the KLD has been proposed as a measure of similarity for various compression and signal processing problems [59, 42, 86, 41].

Given a density $P_1(\mathbf{x})$ and a family of densities $\mathcal{M}$ the minimum discrimination information criteria [83] seeks the density in $\mathcal{M}$ that is the "nearest neighbor" of $P_1(\mathbf{x})$ in the KLD sense

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

If $\mathcal{M}$ is a large family, containing $P_1(\mathbf{x})$, this problem has the trivial solution $P_2(\mathbf{x}) = P_1(\mathbf{x})$, which is not always the most interesting. In other cases, a sample from $P_2(\mathbf{x})$ is available but the explicit form of the distribution is not known. In these situations it may be more useful to seek for the distribution that minimizes the KLD subject to a stricter set of constraints. Kullback suggested the problem

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

subject to

$$\int T(\mathbf{x}) P_2(\mathbf{x}) = \theta$$

where $T(\mathbf{x})$ is a measurable statistic (e.g. the mean when $T(\mathbf{x}) = \mathbf{x}$) and θ can be computed from a sample (e.g. the sample mean). He showed that the minimum is 1) achieved by

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

where Z is a normalizing constant, $Z = \int e^{-\lambda T(\mathbf{x})} P_1(\mathbf{x}) d\mathbf{x}$, and λ a Lagrange multiplier [11] that weighs the importance of the constraint; and 2) equal to

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

Gray and his colleagues have studied extensively the case in which $P_1(\mathbf{x})$ belongs to the family of *auto-regressive moving average* (ARMA) processes [59, 42] and showed, among other things, that in this case the optimal solution is a variation of the Itakura-Saito distance commonly used in speech analysis and compression. Kupperman [84, 83] 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 [39]), the constrained version of MDI is equivalent to maximum likelihood.

The KLD has only been recently considered in the retrieval literature [192, 189, 70, 139, 16], where attention has focused on the unconstrained MDI problem

$$g(\mathbf{x}) = \arg\min_i KL[P_{\mathbf{X}}(\mathbf{x})||P_{\mathbf{X}|Y}(\mathbf{x}|i)], \tag{2.18}$$

where $P_{\mathbf{X}}(\mathbf{x})$ is the density of the query and $P_{\mathbf{X}|Y}(\mathbf{x}|i)$ that of the i^{th} image class. Similarly to the constrained case, it is possible to derive a connection between unconstrained MDI and maximum likelihood. However, the connection is much stronger in the unconstrained case since there is no need to make any assumptions regarding the type of densities involved. In particular, by simple application of the law of large numbers to (2.16),

$$\begin{aligned}
g(\mathbf{x}) &= \arg\max_i E_{\mathbf{x}}[\log P_{\mathbf{X}|Y}(\mathbf{x}|i)] \quad \text{as } N \to \infty \\
&= \arg\max_i \int P_{\mathbf{X}}(\mathbf{x}) \log P_{\mathbf{X}|Y}(\mathbf{x}|i) d\mathbf{x} \\
&= \arg\min_i \int P_{\mathbf{X}}(\mathbf{x}) \log P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} - \int P_{\mathbf{X}}(\mathbf{x}) \log P_{\mathbf{X}|Y}(\mathbf{x}|i) d\mathbf{x} \\
&= \arg\min_i \int P_{\mathbf{X}}(\mathbf{x}) \log \frac{P_{\mathbf{X}}(\mathbf{x})}{P_{\mathbf{X}|Y}(\mathbf{x}|i)} d\mathbf{x} \\
&= \arg\min_i KL[P_{\mathbf{X}}(\mathbf{x})||P_{\mathbf{X}|Y}(\mathbf{x}|i)],
\end{aligned}$$

where $E_{\mathbf{x}}$ is the expectation with respect to the query density $P_{\mathbf{X}}(\mathbf{x})$. This means that, independently of the type of densities, MDI is simply the asymptotic limit of the ML criteria as the cardinality of the query tends to infinity[3]. This relationship is important for various reasons. First, it confirms that the Bayesian criteria converges to a meaningful global similarity function as the cardinality of the query grows. Second, it makes it clear that while ML and MDI perform equally well for global queries, the Bayesian criteria has the added advantage of also enabling local queries. Third, while the Bayesian criteria has complexity $O(N)$, as we will see in Chapter 6, for most densities of practical interest MDI either has a much reduced complexity or can be approximated by functions with that property. In practice, by switching to MDI when the size of the query exceeds a given

[3] Notice that this result only holds when the true distribution is that of the query. The alternative version of the divergence, where the distribution of the database image class is assumed to be true, does not have an interpretation as the asymptotic limit of a local metric of similarity.

threshold, this allows the complexity of Bayesian retrieval to always remain manageable. Finally, it establishes a connection between the Bayesian criteria and several similarity functions that can be derived from MDI.

2.3.4 χ^2 test

The first of such similarity functions is the χ^2 statistic. Using a first order Taylor series approximation for the logarithmic function about $x = 1$, $\log(x) \approx x - 1$, we obtain[4]

$$\begin{aligned}
KL[P_1(\mathbf{x})||P_2(\mathbf{x})] &= \int P_1(\mathbf{x}) \log \frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} d\mathbf{x} \\
&\approx \int \frac{P_1(\mathbf{x})^2 - P_1(\mathbf{x})P_2(\mathbf{x})}{P_2(\mathbf{x})} d\mathbf{x} \\
&= \int \left(\frac{P_1(\mathbf{x})^2 - P_1(\mathbf{x})P_2(\mathbf{x})}{P_2(\mathbf{x})} - P_1(\mathbf{x}) + P_2(\mathbf{x}) \right) d\mathbf{x} \\
&= \int \frac{(P_1(\mathbf{x}) - P_2(\mathbf{x}))^2}{P_2(\mathbf{x})} d\mathbf{x},
\end{aligned}$$

where we have used the fact that $\int P_i(\mathbf{x})d\mathbf{x} = 1, i = 1, 2$. In the retrieval context, this means that MDI can be approximated by

$$g(\mathbf{x}) \approx \arg\min_i \int \frac{(P_{\mathbf{X}}(\mathbf{x}) - P_{\mathbf{X}|Y}(\mathbf{x}|i))^2}{P_{\mathbf{X}|Y}(\mathbf{x}|i)} d\mathbf{x}. \tag{2.19}$$

The integral on the right is known as the χ^2 statistic and the resulting criteria a χ^2 test [124]. It has been proposed as a metric for image similarity in [157, 16, 139]. Since it results from the linearization of the KLD, it can be seen as an approximation to the asymptotic limit of the ML criteria. Obviously, this linearization can discard a significant amount of information and there is, in general, no reason to believe that it should perform better than Bayesian retrieval.

2.3.5 The Gaussian case

Several similarity functions of practical interest can be derived from the Bayesian retrieval criteria when the class likelihoods are assumed to be Gaussian. We now analyze the relationships for three such functions: the quadratic, Mahalanobis, and Euclidean distances.

[4] This result is stated without proof in [31].

Given the asymptotic convergence of ML to MDI, these results could also been derived from the expression for the KLD between two Gaussians [83], by replacing expectations with respect to the query distribution by sample means.

Quadratic distance

When the image features are Gaussian distributed, (2.16) becomes

$$\begin{aligned} g(\mathbf{x}) &= \arg\min_i \log|\mathbf{\Sigma}_i| + \frac{1}{N}\sum_n (\mathbf{x}_n - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mu_i) \\ &= \arg\min_i \log|\mathbf{\Sigma}_i| + \hat{\mathcal{L}}_i, \end{aligned} \tag{2.20}$$

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* (QD) commonly found in the perceptually weighted compression literature [53, 89, 119, 92]. As a retrieval metric, the QD can thus be seen as the result of imposing two stringent restrictions on the generic ML criteria. First, that all image sources are Gaussian and, second, that their covariance matrices are orthonormal ($|\mathbf{\Sigma}_i| = 1, \forall i$).

Mahalanobis distance

Furthermore, because

$$\begin{aligned} \hat{\mathcal{L}}_i = \frac{1}{N}&\sum_n (\mathbf{x}_n - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mu_i) \\ &= \frac{1}{N}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}} + \hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \hat{\mathbf{x}} + \hat{\mathbf{x}} - \mu_i) \\ &= \frac{1}{N}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}})^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \hat{\mathbf{x}}) - 2(\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}\frac{1}{N}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}}) + (\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\hat{\mathbf{x}} - \mu_i)^T \\ &= \frac{1}{N} trace[\mathbf{\Sigma}_i^{-1}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}})(\mathbf{x}_n - \hat{\mathbf{x}})^T] + (\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\hat{\mathbf{x}} - \mu_i)^T \\ &= trace[\mathbf{\Sigma}_i^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}}] + (\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\hat{\mathbf{x}} - \mu_i)^T \\ &= trace[\mathbf{\Sigma}_i^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}}] + \mathcal{M}_i, \end{aligned} \tag{2.21}$$

where

$$\hat{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$$

is the sample mean of $\mathbf{x}_n$

$$\hat{\mathbf{\Sigma}}_{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \hat{\mathbf{x}})(\mathbf{x}_n - \hat{\mathbf{x}})^T$$

the sample covariance and

$$\mathcal{M}_i = (\hat{\mathbf{x}} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\hat{\mathbf{x}} - \mu_i)^T$$

the Mahalanobis distance, we see that the MD results from complementing Gaussianity with the assumption that all classes have the same covariance ($\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{\Sigma}_i = \mathbf{\Sigma}, \forall i$).

Euclidean distance

Finally, if this covariance is the identity ($\mathbf{\Sigma} = \mathbf{I}$), we obtain the square of the Euclidean distance (ED) or *mean squared error*

$$\mathcal{E}_i = (\hat{\mathbf{x}} - \mu_i)^T(\hat{\mathbf{x}} - \mu_i). \tag{2.22}$$

The MD, the ED, and variations on both, have been widely used in the retrieval literature [163, 24, 96, 43, 166, 153, 118, 158, 134, 102, 193, 129, 65, 15, 160, 139, 72, 195, 175, 150, 96, 7].

Some intuition for the advantages of Bayesian retrieval

The Gaussian case is a good example of why, even if minimization of error probability is not considered to be the right goal for an image retrieval system, there seems to be little justification to rely on any criteria for image similarity other than the Bayesian. Recall that, under Bayesian retrieval, the similarity function is

$$g(\mathbf{x}) = \arg\min_i \log|\mathbf{\Sigma}_i| + \overbrace{trace[\mathbf{\Sigma}_i^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}}] + \underbrace{(\hat{\mathbf{x}} - \mu_i)^T\mathbf{\Sigma}_i^{-1}(\hat{\mathbf{x}} - \mu_i)^T}_{\text{MD}}}^{\text{QD}} \tag{2.23}$$

and all three other criteria are approximations that arbitrarily discard covariance information.

As illustrated by Figure 2.2, this information is important for the detection of subtle variations such as rotation and scaling in feature space. 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 $\pi/2$.

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 example, 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.

2.3.6 L^p norms

Despite all the nice properties discussed above, probabilistic retrieval has received small attention in the context of CBIR. An overwhelmingly more popular metric of global similarity is the L^p norm of the difference between densities

$$g(\mathbf{X}) = \arg\min_i \left(\int_{\mathcal{F}} |P_{\mathbf{X}}(\mathbf{x}) - P_{\mathbf{X}|Y}(\mathbf{x}|i)|^p d\mathbf{x} \right)^{\frac{1}{p}}. \tag{2.24}$$

These norms are particularly common in the color-based retrieval literature as metrics of similarity between color histograms.

The *histogram* of a collection of feature vectors $\mathbf{X}$ is a vector $\mathbf{f} = \{f_1, \ldots, f_R\}$ associated with a partition of the feature space $\mathcal{X}$ into R regions $\{\mathcal{X}_1, \ldots, \mathcal{X}_R\}$, where f_r is the number of vectors in $\mathbf{X}$ landing on cell $\mathcal{X}_r$. Assuming a feature space of dimension n and rectangular cells of size $h_1 \times \ldots \times h_n$, the histogram provides an estimate of the feature probability density of the form

$$P(\mathbf{X}) = \sum_k \frac{f_k}{F} \mathcal{K}(\mathbf{x} - \mathbf{c}_k), \tag{2.25}$$

where $\mathbf{c}_k$ is the central point of the k^{th} cell, F the total number of feature vectors, and $\mathcal{K}(\mathbf{x})$

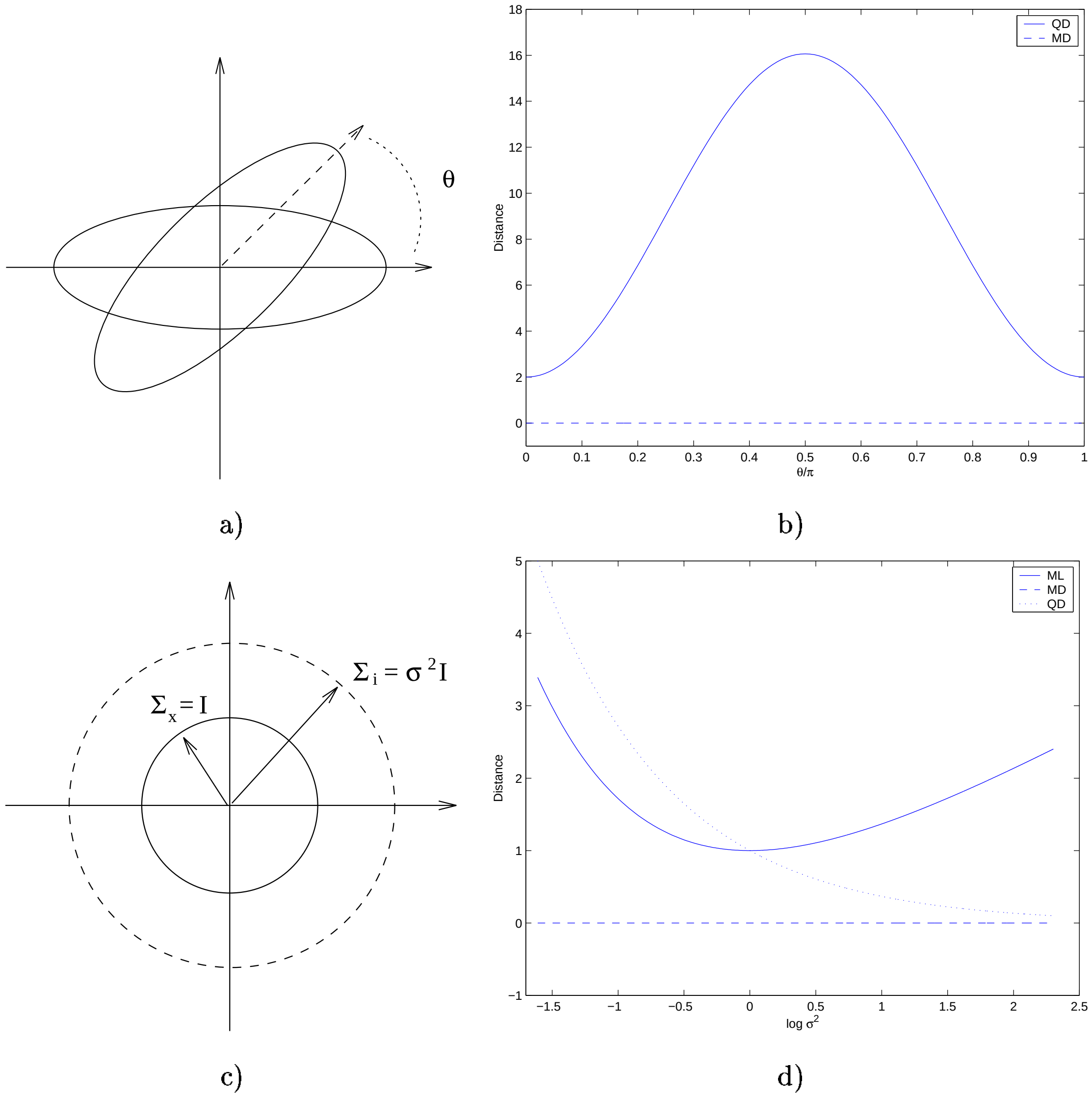

Figure 2.2: a) A Gaussian with mean $(0,0)^T$ and covariance $diag(4, 0.25)$ and its replica rotated by θ. b) Distance between the Gaussian and its rotated replicas as a function of θ/π under both the QD and the MD. c) Two Gaussians with different scales. d) Distance between them as a function of $\log \sigma^2$ under ML, QD, and MD.

a pdf such that

$$\begin{aligned}
&\mathcal{K}(\mathbf{x}) > 0, \;\text{ if } |\mathbf{x}_1| < \frac{h_1}{2}, \ldots, |\mathbf{x}_n| < \frac{h_n}{2}, \\
&\mathcal{K}(\mathbf{x}) = 0, \;\text{ otherwise}, \\
&\int \mathcal{K}(\mathbf{x}) d\mathbf{x} = 1.
\end{aligned}$$

Defining $\mathbf{q}$ to be the histogram of Q query vectors, $\mathbf{p}^i$ the histogram of P^i vectors from the i^{th} image class, and substituting (2.25) into (2.24)

$$\begin{aligned}
g(\mathbf{X}) &= \arg\min_i \left(\int_{\mathcal{X}} |\sum_r \left(\frac{q_r}{Q} - \frac{p_r^i}{P^i} \right) \mathcal{K}(\mathbf{x} - \mathbf{c}_r)|^p d\mathbf{x} \right)^{\frac{1}{p}} \\
&= \arg\min_i \left(\int_{\mathcal{X}} \sum_r \left| \frac{q_r}{Q} - \frac{p_r^i}{P^i} \right|^p \mathcal{K}(\mathbf{x} - \mathbf{c}_r) d\mathbf{x} \right)^{\frac{1}{p}} \\
&= \arg\min_i \left(\sum_r \left| \frac{q_r}{Q} - \frac{p_r^i}{P^i} \right|^p \int_{\mathcal{X}_r} \mathcal{K}(\mathbf{x} - \mathbf{c}_r) d\mathbf{x} \right)^{\frac{1}{p}} \\
&= \arg\min_i \left(\sum_r \left| \frac{q_r}{Q} - \frac{p_r^i}{P^i} \right|^p \right)^{\frac{1}{p}}, \qquad (2.26)
\end{aligned}$$

where we have used the fact that the cells $\mathcal{X}_r$ are disjoint and $\mathcal{K}(\mathbf{x})$ integrates to one. As shown in [172], assuming that the histograms are normalized ($\sum_r q_r/Q = \sum_r p_r^i/P^i = 1, \forall i$), the minimization of the L^1 distance is equivalent to the maximization of the *histogram intersection* (HI)

$$g(\mathbf{X}) = \arg\max_i \frac{\sum_r \min(q_r, p_r^i)}{Q}, \qquad (2.27)$$

a similarity function that has become the de-facto standard for color-based retrieval [172, 139, 149, 96, 72, 150, 163, 164, 125, 68, 44, 167, 43, 168, 17].

It is clear that, while (2.16) minimizes the classification error, (2.24) implies that minimizing pointwise similarity between density estimates should be the ultimate retrieval criteria. Clearly, for any of the two criteria to work, it is necessary that the estimates be close to the true densities. However, it is known (e.g. see Theorem 6.5 of [38]) that the probability of error of rules of the type of (2.16) tends to the Bayes error orders of magnitude faster than the associated density estimates tend to the right distributions. This implies that accurate density estimates are not required everywhere for the classification criteria to work.

In fact, accuracy is required only in the regions near the boundaries between the different classes, because these are the only regions that matter for the classification decisions. On the other hand, the criteria of (2.24) is clearly dependent on the quality of the density estimates all over $\mathcal{X}$. It, therefore, places a much more stringent requirement on the quality of these estimates and, since density estimation is know to be a difficult problem [184, 162], there seems to be no reason to believe that it is a better retrieval criteria than (2.16). We next validate these theoretical claims through retrieval experiments on real image databases.

2.4 Experimental evaluation

A series of retrieval experiments was conducted to evaluate the performance of the ML criteria as a global similarity function. Since implementing all the similarity functions discussed above was an extensive amount of work, we selected the two most popular representatives: the Mahalanobis distance for texture-based and the histogram intersection for color-based retrieval. In order to isolate the contribution of the similarity function from those of the features and the feature representation, the comparison was performed with the feature sets and representations that are commonly used for each of the domains: color-based retrieval was implemented by combining the color histogram with (2.16) and texture-based retrieval by the combination of the features derived from the *multi-resolution simultaneous auto-regressive* (MRSAR) model[5] [104] with (2.23).

The MRSAR features were computed using a window of size 21 × 21 sliding over the image with increments of two pixels in both the horizontal and vertical dimensions. Each feature vector consists of 4 SAR parameters plus the error of the fit achieved by the SAR model at three resolutions, in a total of 15 dimensions. This is a standard implementation of this model [104, 94, 102]. For color histogramming, the 3D YBR color space was quantized by finding the bounding box for all the points in the query and retrieval databases and then dividing each axis in b bins. This leads to b^3 cells. Experiments were performed with different values of b.

Figure 2.3 presents precision/recall curves for the Brodatz and Columbia databases. As

[5] See the appendix for a more detailed justification for the use of the MRSAR features as a benchmark.

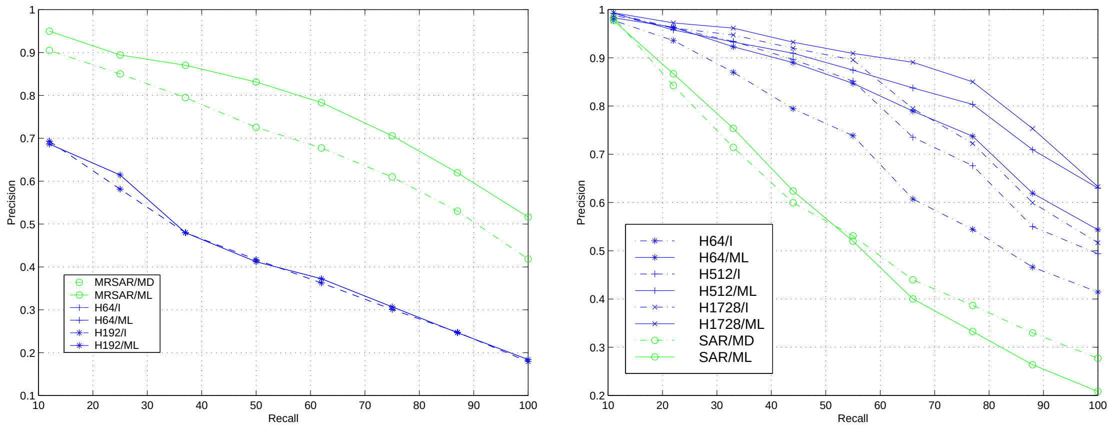

Figure 2.3: Precision/recall curves for Brodatz (left) and Columbia (right). In the legend, MRSAR means MRSAR features, H color histograms, ML maximum likelihood, MD Mahalanobis distance, and I intersection. The total number of bins in each histogram is indicated after the H.

expected, texture-based retrieval (MRSAR/MD) performs better on Brodatz while color-based retrieval (color histogramming) does better on Columbia. Furthermore, due to their lack of spatial support, histograms do poorly on Brodatz while, being a model specific for texture, MRSAR does poorly on Columbia[6].

More informative is the fact that, when the correct features and representation are used for the specific database, the ML criteria always leads to a clear improvement in retrieval performance. In particular, for the texture database, combining ML with the MRSAR features and the Gaussian representation leads to an improvement in precision from 5 to 10% (depending on the level of recall) over that achievable with the Mahalanobis distance. Similarly, on Columbia, replacing histogram intersection by the ML criteria leads to an improvement that can be as high as 20%[7].

The latter observation validates the arguments of section 2.3.6, where we saw that, while

[6] Notice that this would not be evident if we were only looking at classification accuracy, i.e. the percentage of retrievals for which the first match is from the correct class.

[7] Notice that, for these databases, 100% recall means retrieving the 8 or 9 images in the same class as the query, and it is important to achieve high precision at this level. This may not be the case for databases with hundreds of images in each class, since it is unlikely that users may want to look at that many images.

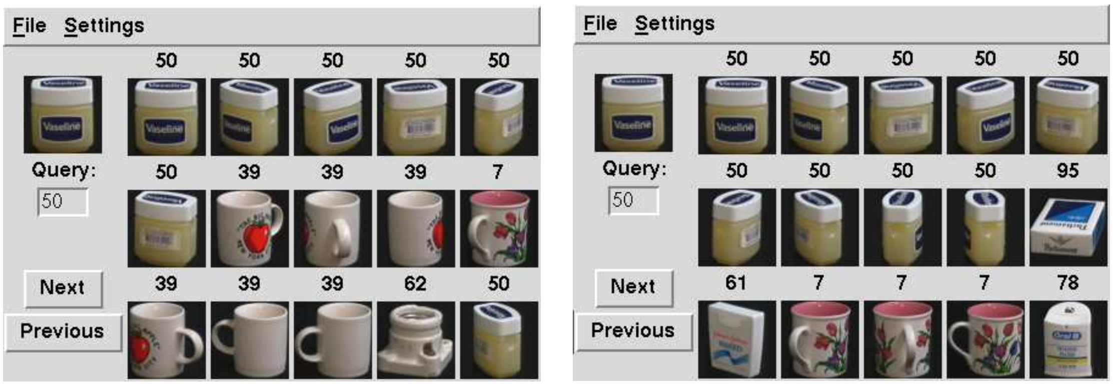

Figure 2.4: Results for the same query under HI (left) and ML (right). In both images, the query is shown in the top left corner, and the returned images in raster-scan order (left to right, top to bottom) according to their similarity rank. The numbers displayed above the retrieved images indicate the class to which they belong.

the ML criteria only depends on the class boundaries, HI measures pointwise distances between densities. This means that whenever there is a change in the imaging parameters (lighting, shadows, object rotation, etc) and the densities change slightly, the impact on HI will be higher than on ML. An example is given in Figure 2.4 where we present the results of the same query under the two similarity criteria. Notice that as the object is rotated, the relative percentages of the different colors in the image change. HI changes accordingly and, when the degree of rotation is significant, views of other objects are preferred. On the other hand, because the color of each individual pixel is always better explained by the density of the rotated object than by those of other objects, ML achieves a perfect retrieval. This increased invariance to changes in imaging conditions explains why, for large recall, the precision of ML is consistently and significantly higher than that of HI.