On the Choice of Similarity Measures for Image Retrieval by Example

Jean-Philippe Tarel and Sabri Boughorbel
INRIA, Rocquencourt
Domaine de Voluceau, BP-105
F-78153 Le Chesnay Cedex, France
Jean-Philippe.Tarel@inria.fr Sabri.Boughorbel@inria.fr

ABSTRACT

In image retrieval systems, a variety of simple similarity measures are used. The choice for one similarity measure or another is generally driven by an experimental comparison on a labeled database. The drawback of such an approach is that, while a large number of possible similarity measures can be tested, we do not know how to extend from the obtained results. However, the choice of a good similarity measure leads to noticeable better results. It is known that this choice is related to the variability of the images within the same class. Therefore, we propose a model of image retrieval systems and deduce a scheme for deriving the best similarity measure in a set of similarity measures, assuming a parametric model of the variability of feature vectors within the same class. An experimental validation of the model and the derived similarity measures is performed on synthetic ground-truth databases. Finally, from our experiments, we give several rules to follow for the design of ground-truth databases allowing reliable conclusions on the search of better similarity measures.

1. INTRODUCTION

In image retrieval systems by example, as in many other computer vision systems, the information of interest in each image is summarized in the so-called signature. The aim of introducing such a summary is to reduce the amount of information to be processed.

An image is thus represented by a feature vector lying in a high dimensional space. Many feature spaces were proposed and experimentally compared. We believe that an analysis of how those feature spaces are compared is required to progress in building better image retrieval systems. Of acourse, this analysis is not easy to perform, and therefore we focus on comparing feature spaces representing in essence the same information about images, but in different ways. In such a case, if quantizing errors are ignored, two feature spaces are equivalent if the associated similarity measures are correctly chosen. As a consequence, the comparison of feature spaces containing essentially the same information leads to the question: what is the best similarity measure that must be used to compare two feature vectors within the same feature space?

In the context of image indexing, many similarity measures were proposed, see [6, 8] for a summary. Two different similarity measures used on the same feature space are usually compared in terms of precision-recall diagram on an image database. But as noticed in [5, 7, 1], the choice of similarity measure is mainly related to the variability of the feature vectors. Thus, we based our approach on a statistical analysis. Given a set of possible similarity measures, we propose a scheme for deriving the best similarity measure in this set, from a parametric model of the variability of feature vectors.

In Section 2, we experimentally show improvements that can be expected by well adapting the similarity measure to the feature space. Experiments are mainly performed on a feature space of color histograms, but other kind of feature spaces are also used. The analysis of these results requires a model of image retrieval systems. In Section 3, we present our model of image retrieval systems. This model gives the criterion a similarity measure must maximize. Then in Section 4, we propose a scheme to derive the best similarity measure that must be used in a set of similarity measures, given the statistical model of the used features. In the next section, we investigate possible parametric models for color histogram features from an experimental point of view. In Section 6 and 7, we derive the best similarity measures for well-known pdfs, such as Gaussian, uniform and exponential, for different sets of similarity measures. We take advantage of these derivations to compare the model to simulations on synthetic experiments. Then, we give some clues on which specifications a ground-truth database must follow to obtain reliable conclusions that can be used for other databases.

2. SIMILARITY MEASURE OPTIMIZATION

It is well known that given a feature space, results of image retrieval may be improved just by a better choice of the similarity measure. In order to run systematic experiments, we use ground-truth image databases and all the results are averaged over all query images. Fig. 1 shows a few image classes from ground-truth database DB_s. Usually, a ground-truth image database is composed of N_c classes with different visual contents. Each class contains N_i images, so the total number of images in the database is N_cN_i. For any query q took from the database, we select the N_i closest images. When the recognition is perfect, the closest images are all from the class to which the query q belongs.

Figure 1: Few images from our database DB_s. It is formed by $N_c = 64$ classes with $N_i = 9$ images per class.

Different ground-truth databases have been used for our experimentations:

- Scene database DB_{sc}: It consists of 81 classes of 9 images each (729 images), and contains TV broadcasts, sample images of videos, paintings...
- Texture Database DB_t: It groups 792 images into 88 classes of 9 images each. It is extracted from Brodatz color texture Database. Color variability in every class is less important than in DB_{sc}.
- Shape Database DB_s: It is formed by 64 classes with 9 images per class (576 images).
- Fit Database DB_f: It contains 7 classes, 100 images per class (700 images). This database is used only for pdf fitting as described in Section 5.
- Corel Database DB_c: It contains 71 classes, 30 images per class (2130 images). It is extracted from the Corel database.

Precision-recall diagrams are widely used to evaluate image retrieval systems. Nevertheless, we decided to use only the average precision to compare different similarity measures. We defined the average precision to be the average of the number of relevant image among the N_i retrieved images, over all the query images of the database.

Without any other knowledge, the best similarity measure has to be searched with a brute force approach. The computational cost of this approach implies that we must reduce the search to a small set of similarity measures. We have reduce ourselves to the set of similarity measures with the following two parameters:

$$S_{\alpha,\beta}(q,s) = \sum_{i=1}^{N_f} (q_i^\alpha - s_i^\alpha)^\beta \qquad (1)$$

where $q = (q_i)$ is the query vector and $s = (s_i)$ is the feature vector to be compared with the query. N_f is the dimension of feature vectors. This kind of similarity measures was previously used with advantages by [2] on Corel image database with color histograms. Indeed the used kernel is of the form $K_{\alpha,\beta}(q,s) = e^{-S_{\alpha,\beta}(q,s)}$. Notice that when $\alpha = 1$ and $\beta > 1$, the similarity measure $S_{1,\beta}(q,s)$ reduces to the L_β distance.

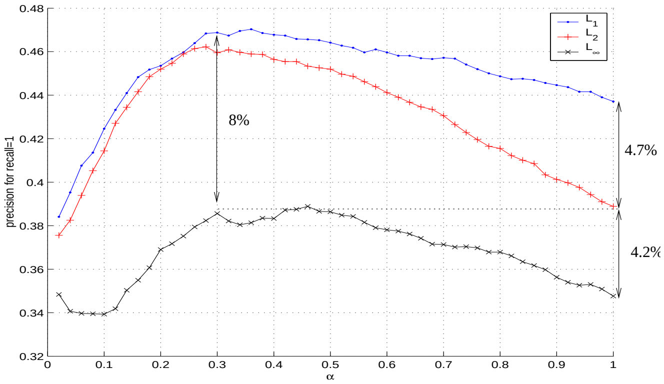

Figure 2: α-diagrams with classical L_1, L_2 and L_∞ distances. α-diagram plots the average precision versus α, using similarity measure $S_{\alpha,\beta}$, on DB_s.

It is well known that the choice of a good β can improve image retrieval system results. For instance, as shown on Fig. 2 for $\alpha = 1$, the use of L_2 rather L_∞ improves the results by 4% in average precision. Moreover, the use of L_1 rather than L_2 improves the results by an extra 5%. Similar results were observed on several other ground-truth databases.

Less known is the importance of a good choice of parameter α. We define α-diagram as the curve of average precision when we vary parameter α in similarity measure $S_{\alpha,\beta}$ in (1), for a given β. Fig. 2 gives a comparison between α-diagrams for $\beta = 1$, $\beta = 2$ and $\beta = +\infty$. Tests have been performed on DB_s. The signature used is simply the RGB color histogram. The obtained curves present a maximum around 0.3. We performed tests on the other ground-truth databases DB_{sc}, DB_t and DB_c, and we obtained the same kind of shapes with RGB color histograms. Moreover, the position of the maximum varies a little depending of the database. Therefore the α-diagram seems mainly related only to the used feature space. Notice that an adequate α ($\alpha = 0.3$ with $\beta = 1$) improves the average precision by more than 8% compared to the Euclidean distance $S_{1,2}$.

Then, we checked that the similarity measure can be optimized for other feature spaces. Fig. 3 shows α-diagrams for: Edge Orientation Histogram (eoh) [3] for shape features, Fourier descriptors for texture features, RGB color histogram and Laplacian weighted RGB color histogram [9] for color features. In this experiment $\beta = 1$, and thus, when $\alpha = 1$, the used similarity measured is the classical L_1 dis-

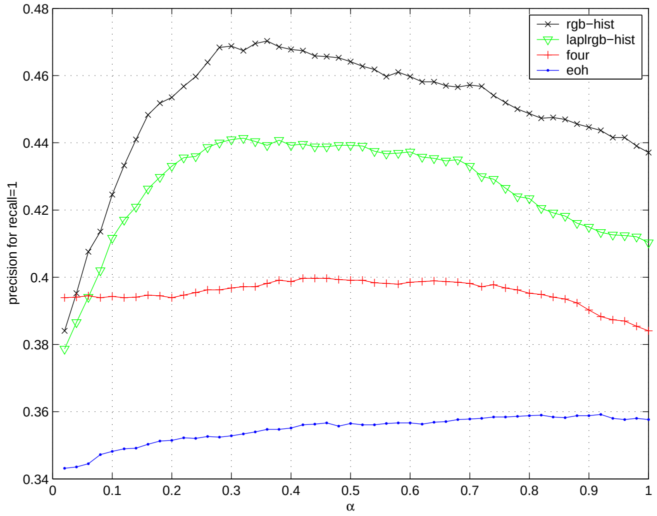

Figure 3: α-diagrams for different feature spaces and $\beta = 1$. Used representations are: Edge Orientation Histogram (eoh) for shape, Fourier descriptors for texture, RGB color and Laplacian weighted RGB color histograms.

tance. By tuning the parameter α, we are able to improve the average precision up to 3% for RGB color histogram and more than 1% for Fourier descriptors, for instance.

From our few user experiments, an increase of 2% of average precision is noticeable. In particular the ordering of the retrieved images seems improved. This indicates, how important is the search of better similarity measures.

There are two main problems with the experimental approach described before. First, it is not possible to perform experiments on all the possible similarity measures that one can imagine, the search space is too large. Second the best similarity measure is obtained only for small ground-truth databases, and we have no clue how this result generalizes to larger databases.

The alternative to the experimental approach is to model image retrieval systems to try to progress in the solution of these two problems. From a model of the observed perturbations of the feature space, we show next how to derive the best associated similarity measure within a set of similarity measures.

3. STATISTICAL MODEL

Contrary to [7], we believe that image retrieval by example should not be modeled as a classification problem. Indeed, classifying a database allows image retrieval, but does not perform as well as classical image retrieval techniques. From a classification point of view, image retrieval is seen as finding a mapping from the database images to the set of class labels. Therefore, if the image retrieval system is perfect, given an image drawn from class C, the retrieved images are all within C.

A key advantage of image retrieval by example is, that the retrieved images from two queries are generally different, even if these queries are in the same class. But, this property is not verified when the image retrieval by example is formulated as a classification problem because the class of the query is not known. We prefer, therefore, to better analyze what is usually performed in image retrieval systems by example.

Usually, visual image information is represented by a feature vector, in order to increase speed and to save memory. Given a feature vector query q, we define the goal of a content-based retrieval system to retrieve the feature vectors that are the most likely to be in the same class as the query. The main difference, compared with classification approach, is, that we do not know the query class. We just know that q is a random realization of its class. Then, given any feature vector s in the database, the problem is to decide if s and q are in the same class or not. If the number of samples is large enough, this question can be answered by statistical testing. Unfortunately, in the context of image retrieval, we only have two samples: s and q.

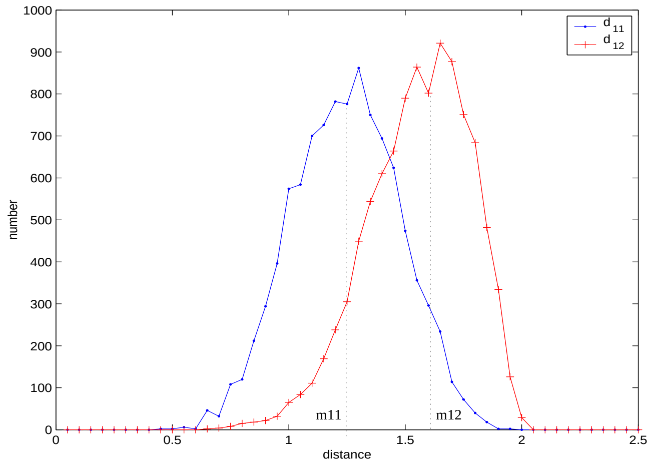

Figure 4: Example of intra-distance d_{11} and inter-distance d_{12} distributions for RGB feature vector from the two classes displayed in Fig. 6 (castle and sunset from the Corel database).

For sake of simplicity, we suppose that only two different classes C_1 and C_2 are available in the database and that q belongs to C_1. The database is also assumed ground-truth. When s also belongs to C_1, the similarity measure $u = d(q, s)$ has a probability distribution function (pdf) $d_{11}(u)$. When s belongs to C_2, a different pdf $d_{12}(u)$ is obtained for u as shown in Fig. 4 on a real two classes database. The distance $u = d(q, s)$ between q and s is always positive and thus pdfs d_{11} and d_{12} are defined only on $[0, +\infty]$. Usually, an image retrieval system is equivalent to a sort on the set of distances of the query to each image of the database. Therefore, given a distance threshold t, the number of images s of C_1 with a lower distance than t to q, is $N_i D_{11}(t) = N_i \int_0^t d_{11}(u)du$. D_{11} is by definition the cumulative distribution function (cdf) of d_{11}. Similarly, the number of images s of C_2 with a lower distance than t to q, is $N_i D_{12}(t) = N_i \int_0^t d_{12}(u)du$. Thus, the total number of retrieved images is thus $N_i(D_{11}(t) + D_{12}(t))$. The percentage of relevant images is $\overline{P} = \frac{D_{11}(t)}{D_{11}(t)+D_{12}(t)}$. $\overline{P}$ is the average precision we introduced in the previous section. The maximization of average precision is used in the following to compare two different similarity measures on the same feature space.

But rather than a threshold on the distance, image retrieval systems better use a threshold on the number of retrieved images. In our two classes problem, we need to return 50% of the total number of images in the database. If the retrieval system is perfect, all the retrieved images are in C_1. To retrieve N_i images, threshold t must be chosen in such a way that $D_{11}(t)+D_{12}(t) = 1$, and thus the average precision reduces to:

$$\overline{P} = D_{11}((D_{11} + D_{12})^{-1}(1)) \tag{2}$$

where t does not appear anymore. This is illustrated in Fig. 5 with two far apart Gaussian pdfs, where m_{11} and m_{12} are the mean of the inter- and intra-distances, respectively.

We now need to link cdfs D_{11} and D_{12} to the pdfs of the feature vectors. Let $P_q(x) = P(x|C(q))$ be the pdf of feature vector x knowing the class of the query q, and $P_s(x) = P(x|C(s))$ be the pdf of the class of s in the database. These pdfs are defined on the feature space $\mathcal{F}$, i.e x belongs to $\mathcal{F}$.

In practice it is not realistic to assume that these pdfs are fully known. A first difficulty is that $\mathcal{F}$ is usually a feature space of high dimensionality. As a consequence, the direct sampling of these pdfs requires a large number of image samples, not possible to achieve in practice. Thus, the components of the feature vector are assumed independent from each other. This assumption means that $P(x)$ is a product of 1D pdfs, and is thus easier to sample. A second difficulty is that this model is still too complicated for a formal analysis. As a consequence, in the proposed parametric model, the components of the feature vectors are assumed independent and having the same kind of pdf with different parameters. For simple derivations, we also assume that the only different parameter is the mean on each component.

More formally, we assume $q_i - \overline{q_i}$ and $s_i - \overline{s_i}$ are iid with a centered pdf e, where q_i and s_i are the components of q and s respectively. $\overline{x}$ denotes the expectation of a random variable x.

4. OPTIMIZATION CRITERION

In this section, we approximate the average precision $\overline{P}$ in (2) to better analyze its link with the centered pdf e of feature vector components.

Generally, similarity measures proposed in the literature are sums of similarity measures of the components. Thus, the distance $d(q,s)$ can be written as:

$$d(q,s) = \sum_{i=1}^{i=N} \delta(q_i - s_i)$$

Let $f(x_i)$ be the pdf of the random variables $x_i = q_i - s_i - m_i$, with $m_i = \overline{q_i} - \overline{s_i}$. Since $q_i - \overline{q_i}$ and $s_i - \overline{s_i}$ have same pdf e, it is not difficult to prove that $f(x_i)$ is a symmetric function with respect to 0. Indeed, we have:

$$f(x) = \frac{1}{2}\int_y e(\frac{x+y}{2})e(\frac{-x+y}{2})dy \tag{3}$$

with $y_i = q_i - s_i - m_i$. Thus, $v_i = q_i - s_i$ has a centered and symmetric pdf when s is in the same class than q. On the contrary, $v_i = q_i - s_i$ has only a symmetric pdf with respect to m_i.

The pdf of the intra-distance u, for one component i, is $f(\delta^{-1}(u))\delta^{-1'}(u)$. The pdf of the intra-distance is the convolution of $f(\delta^{-1}(u))\delta^{-1'}(u)$ with itself, N_f times. Indeed, the intra-distance d_{11} is the sum on all components of the intra-distance δ_{11}. Using the central limit theorem (and thus assuming finite variance for u [4]), the pdf d_{11} converges towards a Gaussian distribution, when N_f goes to infinity. A similar result can be derived for the inter-distance [4]: the pdf d_{12} converges towards a Gaussian distribution, when N_f goes to infinity. As an illustration, with $N_f = 180$, intra-distance d_{11} and inter-distance d_{12} of Fig. 4 look like Gaussian distributions.

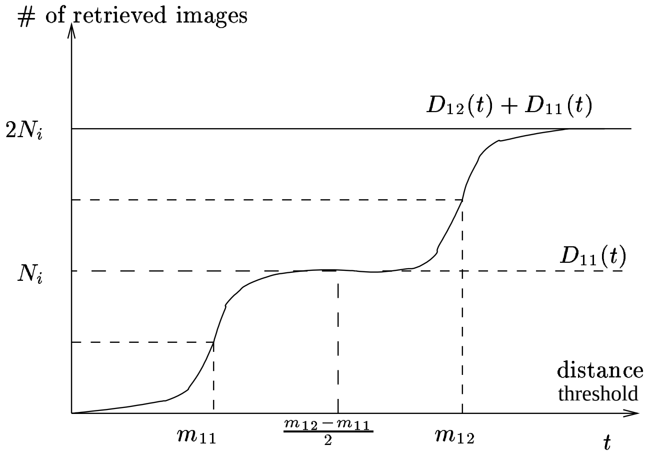

Figure 5: Number of images retrieved from the database as a function of the distance t used as a threshold.

The two pdfs d_{11} and d_{12} being approximatively Gaussian, they are summarized by their mean and variance. The mean of the intra-distance is simply given by $m_{11} = \overline{d_{11}(v)} = N_f\overline{\delta(v)}$. Its variance is $v_{11} = N_f(\overline{\delta^2(v)} - \overline{\delta(v)}^2)$.

The case of inter-distance is a little more complicated. When the different component means m_i are large with respect to the standard deviation of v_i, the two pdfs d_{11} and d_{12} are far apart, and the retrieval system gives perfect results. A more realistic situation is when m_i is small with respect to the standard deviation of v_i. This observation leads us to perform a second order Taylor expansion on $\delta(v_i + m_i)$ with respect to m_i:

$$\delta(v_i + m_i) \simeq \delta(v_i) + m_i\delta'(v_i) + \frac{1}{2}m_i^2\delta''(v_i)$$

With this approximation, we deduce that the mean of inter-distance is $m_{12} = \overline{d_{12}(v)} \simeq N_f\overline{\delta(v)} + \frac{1}{2}\sum_{i=1}^{i=N_f} m_i^2\overline{\delta''(v)}$ since $\delta'(-v) = -\delta'(v)$. With the zero order expansion, the variance of the inter-distance is $v_{12} \simeq N_f(\overline{\delta^2(v)} - \overline{\delta(v)}^2)$. We will see later, that it is sufficient to use order zero for consistent expansion of $\overline{P}$.

The two pdfs d_{11} and d_{12} being approximatively Gaussian, it also allows us to derive a simpler equation for the average precision $\overline{P}$ in (2). We denote as $g(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$ the reduced Gaussian pdf, and $G(x) = \int_{-\infty}^{x} g(t)dt$ its cdf. $G(x)$ is also known as the error function. By symmetry of $G(x)$ and since $v_{11} = v_{12}$, as shown in Fig. 5, we deduce that $t = \frac{m_{11}+m_{12}}{2}$ is the t where $G_{11}(\frac{t-m_{11}}{\sqrt{v_{11}}}) + G_{12}(\frac{t-m_{12}}{\sqrt{v_{12}}}) = 1$. From (2), we then deduce:

$$\overline{P} \simeq G(\frac{m_{12} - m_{11}}{2\sqrt{v_{11}}}) \tag{4}$$

From (4), when there is no difference in means ($m_{12} = m_{11}$), the two classes have exactly the same pdf and thus the retrieval system gives a completely random result, i.e $\overline{P} = 0.5$.

Now, we substitute the obtained means and variances of d_{11} and d_{12} in the average precision (4):

$$\overline{P} \simeq G(\frac{\sqrt{N_f}\,\overline{m_i^2}}{2}\frac{\overline{\delta''(v)}}{2\sqrt{\overline{\delta^2(v)} - \overline{\delta(v)}^2}}) \tag{5}$$

This last equation is important since it allows us to predict the average precision obtained by an image retrieval system, knowing the pdf of the difference v of feature vector components, as a function of the component distance δ. This equation opens the possibility of analytical studies of the optimal similarity measure based on the feature vector pdf e, for usual image retrieval systems.

We denote as G^{-1} the inverse of the error function G. Then, we introduce the following rectified average precision:

$$r_\delta(v) = G^{-1}(|\overline{P}|) = \frac{\sqrt{N_f}\,\overline{m_i^2}}{2}\frac{\overline{\delta''(v)}}{2\sqrt{\overline{\delta^2(v)} - \overline{\delta(v)}^2}} \tag{6}$$

The rectified precision $r_\delta(v)$ is always positive. And, the higher $r_\delta(v)$ is, the better is the image retrieval system. A rectified precision of zero means that the system performs very poorly like a random sampler. An infinite rectified precision means that the system performs perfectly.

5. FEATURE PDF FITTING

In practice, the pdf of the feature vector components are usually not known. A question arises: what kind of pdf is useful to model image features used in image retrieval systems? Due to the reduced number of images per class, we have to assume a parametric model with only a few number of unknown parameters.

Using a database built from ground-truth classes of images, the shape of these pdfs for each component can be estimated. However, the number of elements per class should be large enough to allow reliable conclusions.

We tried to choose a wide panel of pdfs to be fitted. We consider the following pdfs:

- One parameter pdfs: Poisson, Exponential.
- Two parameters pdfs: Rayleigh, Weibull, Gamma, Gaussian, Uniform and Lognormal.

In [1], the authors started to investigate statistical approaches for modeling content-based retrieval systems, and they also investigated a parametric fitting of component pdfs. Nevertheless, our approach turns out to be rather different due to the fact that we do not assume that all the parameters are fully known when the retrieval system processes a query. This leads us to search for the best similarity measure rather to transform the pdfs of the components into uniform pdfs.

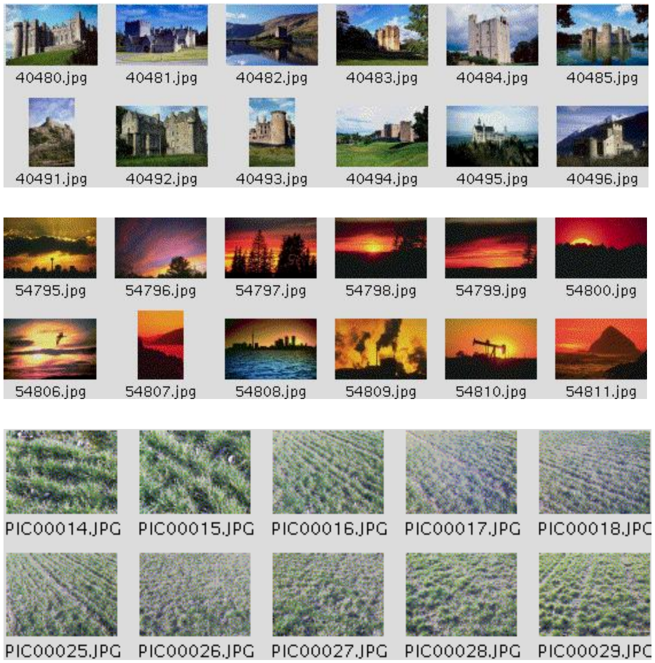

Figure 6: Image examples from three classes: castle, sunset and grass. Notice that the variability in colors from these three classes over 100 images is rather different.

To experimentally test the different models, we performed a Kolmogorov-Smirnov test on 700 images of a ground-truth database. We choose the RGB color histogram as feature vector because, with color, it is easier to build classes. Given a color space quantization, the color histogram is a measure of the color distribution in the image, i.e, each component of the RGB histogram represents the probability that a pixel of the image has the corresponding color. As explained before, we consider that components are independent of one another, and thus we focus on the color pdf of each component separately within an image class. All images in a class contain almost the same visual information, and each class has a different level of color variability, as illustrated in Fig. 6. We used 7 classes of 100 images each: 5 classes from the Corel database (sunset, mountain, waterfall, castle, and raptor), a grass texture class that we collected ourselves, and finally an outdoor class having 100 random images from the Corel database of very different outdoor scenes.

During the tests, we took into account the fact that color

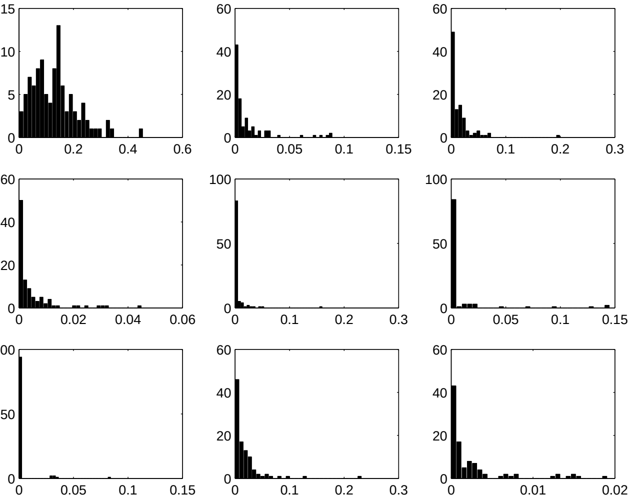

Figure 7: Examples of empirical pdfs of RGB feature vector from the castle class. Each sub-plot displays the number of pixels with the same color, as a percentage of the image size, versus the color.

histograms are sparse (on average more than 70% of the components are null), at least for the color quantization that we choose (6 bins per color axes and thus 216 components for the color histogram). We notice that component pdfs are either mono- or bi-modal. Usually, there is an important mode in 0. Fig. 7 shows typical examples of the empirical pdfs.

With 100 values, it is possible to perform maximum likelihood estimation of the parameters of each pdf. A Kolmogorov-Smirnov test [4] is then used to evaluate the goodness-of-fit. Recall that, this test consists in comparing, with L_∞ distance, the experimental and theoretical cumulative distribution functions. The L_∞ distance is defined as the maximum bin difference. This test provides a criterion for model rejection.

Fig. 8 presents Kolmogorov-Smirnov results performed on the castle and grass classes. Histograms show the number of feature components that may fit each pdf model. It is clear that Exponential pdf performs better than Poisson pdf, for one parameter pdfs. Weibull and gamma pdfs perform better compared to the others with two parameters. Similar results have been observed for the other classes.

Due to the complexity versus fit dilemma, it is difficult to have a definitive conclusion on which pdf the feature components is following. Nevertheless, the following family of pdfs seems suitable:

$$h(x|s,a,b) = \frac{a}{\Gamma(b)s^b} x^{ab-1} e^{-\frac{x^a}{s}} I_{x>0} \qquad (7)$$

obtained by applying the power function $x^{\frac{1}{a}}$ on gamma random variable x. By definition, the Euler function is $\Gamma(b) = \int_{t=0}^{t=+\infty} t^{b-1} e^{-t} dt$. Half-Gauss, Gamma, Weibull, Exponential, Rayleigh pdfs are all particular cases of this family. For instance, the Half-Gaussian is obtained for $a = 2$ and $b = 1/2$.

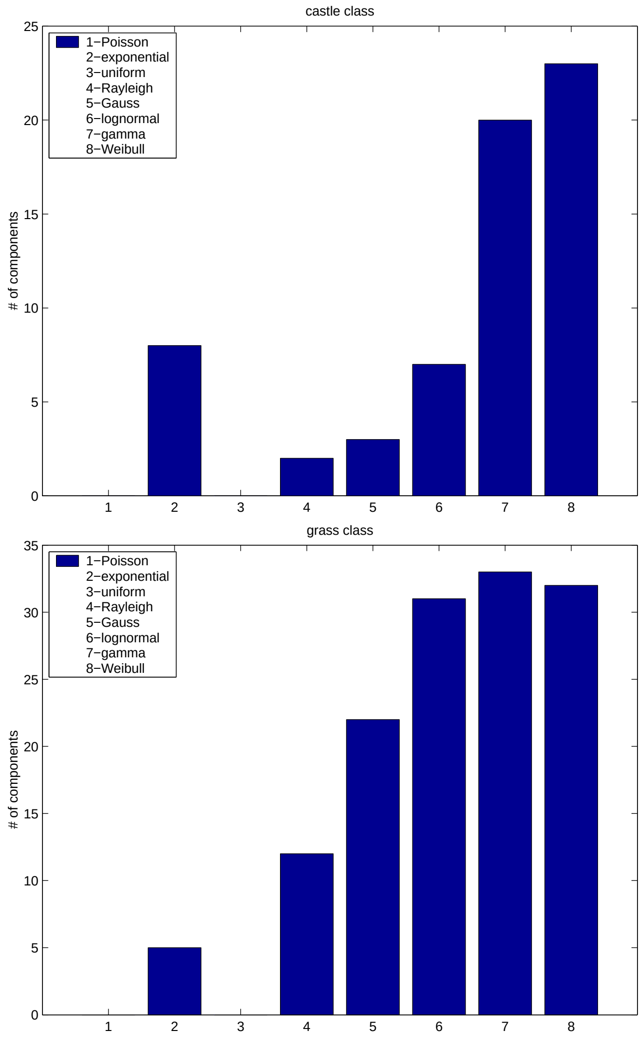

Figure 8: Result of the Kolmogorov-Smirnov test on the castle and grass classes. The number of components not rejected is shown versus the pdf name.

6. OPTIMIZATION OVER β

Compared to (1), in this section, we restrict the set of similarity measures, where we search for the best one, to L_β distances, i.e $\alpha = 1$. Thus, we have $\delta(v) = |v|^\beta$.

With examples, we illustrate how the best similarity measure within L_β can be derived depending of the pdf modeling the variability of the feature components. Then, we take advantage of these derivations to validate the proposed model.

When the similarity measure is L_β, it is interesting to notice that the rectified precision in (6) can be written as the product of two terms:

$$r_\beta(v) = \frac{\sqrt{N_f}\,\overline{m_i^2}}{2\overline{v^2}} \frac{\beta(\beta-1)\overline{|v|^{\beta-2}}\,\overline{v^2}}{2\sqrt{\overline{|v|^{2\beta}} - \overline{|v|^{\beta}}^2}} \qquad (8)$$

Notice that r_β is defined only for $\beta > 1$.

In (8), the first term $\frac{\sqrt{N_f}\,\overline{m_i^2}}{2\overline{v^2}}$ involves only the dimension N_f of the feature space and the average squared difference of component means m_i, relative to the variance $\overline{v^2}$ of v. This term is not related to β, and thus to the similarity measure. We named the second term squared, the speed $\tau_\beta(v)$ of the precision:

$$\tau_\beta(v) = \frac{\beta^2(\beta-1)^2\overline{|v|^{\beta-2}}^2\,\overline{v^2}^2}{4(\overline{|v|^{2\beta}} - \overline{|v|^{\beta}}^2)} \tag{9}$$

The speed $\tau_\beta(v)$ is related to the choice of β and to the shape of the pdf f of the components differences v (and thus to the shape of e). But notice that $\tau_\beta(v)$ is invariant under scale variations of v. This allows us to simplify the following derivations by assuming that the variance of v is 1.

The search for the best similarity measure within L_β is equivalent to maximizing the speed $\tau_\beta(v)$ with respect to β.

6.1 Gaussian

The simplest model is to assume that the N_f components (q_i) of the feature vector q are Gaussian random variables following:

$$g(q|\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{q^2}{\sigma^2}} \tag{10}$$

with mean zero and standard deviation σ. With the previous notations, we have $e(q) = g(q)$.

The components being Gaussian with variance σ^2, a components difference v is also Gaussian with variance $2\sigma^2$. Thus pdf $f(v)$ is also Gaussian. Before to compute the average precision (5), we need the moment of order p of v:

$$\overline{|v|^p} = (2\overline{v^2})^{\frac{p}{2}}\frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{1}{2})}$$

which is defined only when $p > -1$. After substitution in (9), we deduce the following speed, after simplifications:

$$\tau_\beta(Gauss) = \frac{\beta^2}{4}\frac{\Gamma^2(\frac{\beta+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{2\beta+1}{2}) - \Gamma^2(\frac{\beta+1}{2})} \tag{11}$$

Notice that from our derivation, the speed is defined only for $\beta > 1$.

To validate the proposed model of image retrieval systems, we compared the average precision obtained in the Gaussian case (11) to the average precision obtained by simulation on a two classes synthetic database of 200 images each. The number of components N_f is set to 100. The synthetic database is simply produced by sampling a Gaussian random variable with $m_i = 1$ and $\sigma = 4$. Fig. 9 shows the average precision obtained for different values of β in the range $[0, 4]$, compared to theoretical values. The fit between predicted and simulated average precisions is relatively good in such a case since the standard deviation of v is relatively large compared to the value of m_i.

From Fig. 9, we can see that the average precision is maximum for $\beta = 2$. This can be also proved analytically. Therefore, when the feature pdf is Gaussian, L_2 is the best similarity measure within the set L_β.

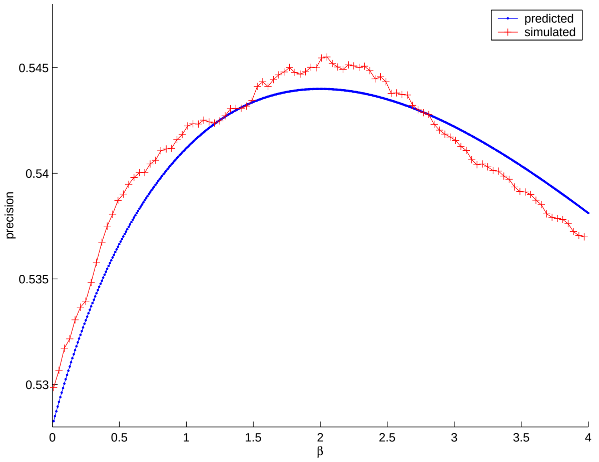

Figure 9: Comparison between theoretical and simulated average precision in the Gaussian case when β varies.

6.2 Exponential

Typically, the feature components are histograms and thus their values are normalized between 0 and 1. Thus, to be exact, previous pdf are typically restricted to $[0, 1]$.

We now assume that feature components are exponential random variables with pdf:

$$\epsilon(q|\sigma) = \frac{1}{\sigma}e^{-\frac{q}{\sigma}}I_{q>0}$$

Its mean is σ and its standard deviation σ. It is simple to prove that the difference of two exponential random variables of parameter σ is a Laplace (or double exponential) random variable with pdf:

$$l(v|\sigma) = \frac{1}{2\sigma}e^{-\frac{|v|}{\sigma}}$$

Its mean is zero and its variance $2\sigma^2$. The moment of order p of v following a Laplace pdf is:

$$\overline{|v|^p} = (\frac{\overline{v^2}}{2})^{\frac{p}{2}}\Gamma(p+1)$$

Therefore, its speed is after simplifications:

$$\tau_\beta(Laplace) = \frac{\Gamma^2(\beta+1)}{\Gamma(2\beta+1) - \Gamma^2(\beta+1)}$$

which is defined for $\beta > 1$ only.

Fig. 10 displays the theoretical and simulated average precisions as a function of β. The simulation is done with the same parameters than in the previous section. The fit between the two curves is not bad for $\beta > 1$, remembering the two approximations we did in Section 4. For $\beta < 1$, out of the domain of definition of speed, $\tau_\beta(Laplace)$ does not nicely extend, like in the Gaussian case. The proposed model is limited a search within L_β with $\beta > 1$.

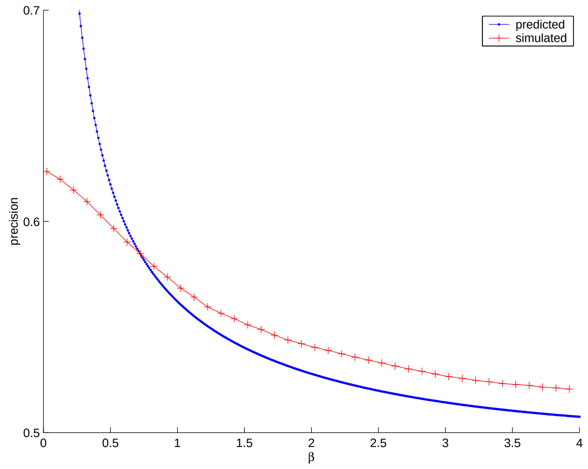

Figure 10: Comparison between theoretical and simulated average precision in the Exponential case when β varies.

As a summary, with the proposed model, we have seen that for an exponential pdf, the optimal similarity measure within L_β is L_1.

6.3 Exponential of a Power

We have previously shown that the best similarity measure within L_β is L_2 when f is a Gaussian pdf and L_1 when f is a Laplace pdf. A question arises: is there a direct link between the power inside the exponential in the pdf and the optimal value of β?

More formally, f is assumed to be:

$$i(v|s,b) = \frac{1}{2b\Gamma(b)s^b}e^{-\frac{|v|^{\frac{1}{b}}}{s}}$$

The mean of v is zero and its variance is $\overline{v^2} = s^{2b}\frac{\Gamma(3b)}{\Gamma(b)}$. Its moment of order p is:

$$\overline{|v|^p} = s^{bp}\frac{\Gamma(b(p+1))}{\Gamma(b)}$$

and thus the speed is:

$$\tau_\beta(i) = \frac{\Gamma^2(3b)}{4\Gamma^2(b)b^2}\frac{\Gamma^2(b(\beta-1)+1)\beta^2}{\Gamma(b)\Gamma(b(2\beta+1))-\Gamma^2(b(\beta+1))}$$

Compared to the previous examples, formal maximization of τ_β with respect to β is rather complicated. Thus, we have numerically search for the curve of the best β, for different values of b.

The obtained results are shown in Fig. 11. Surprisingly, it appears that $L_{\frac{1}{b}}$ is not in general the best similarity measure. For all $\frac{1}{b}$ in $[1, 1.7364]$, L_1 is the best distance within L_β, with $\beta > 1$. This may explain why L_1 shows better results compared to L_2 and $L_{+\infty}$ in the experiments of Fig. 2.

6.4 Uniform

In Fig. 11, when $\frac{1}{b}$ goes not infinity, i.e, when $i(v|s,b)$ becomes the distribution function of a uniform random variable, β seems to go towards infinity. To check this, we compute the speed in the case where the feature components are uniform random variables. In such a case, the difference of two components is a centered triangular random variable (or Simpson [4]) with moments of order p:

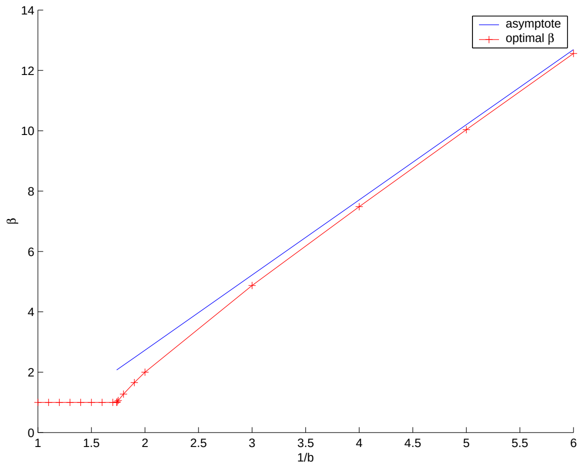

Figure 11: Best β as a function of $\frac{1}{b}$ for a pdf in the family $i(v|s,b)$.

$$\overline{|v|^p} = 2\frac{(6\overline{v^2})^{\frac{p}{2}}}{(p+1)(p+2)}$$

The obtained speed is defined for $\beta > 1$ as:

$$\tau_\beta(triangle) = \frac{(\beta+1)^2(\beta+2)^2(2\beta+1)}{36\beta^2(\beta+5)}$$

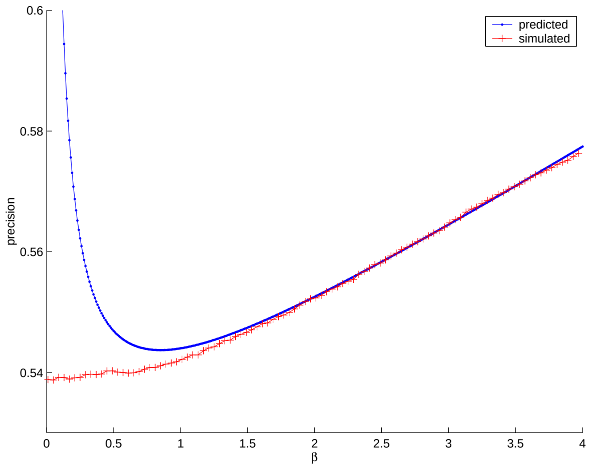

Figure 12: Comparison between predicted and observed average precision in the uniform case when β varies.

Another time, as shown in Fig. 12, the predicted and simulated average precisions are very close to each other for $\beta > 1$. The curve is increasing from $\beta = 1$, and thus L_∞ is the best similarity measure within L_β, $\beta > 1$, for a uniform pdf.

7. OPTIMIZATION OVER α

In the previous section, we optimized the similarity measure over β. But as shown in Section 2, an optimization over α is also of interest. To simplify the analysis, we assume that β is fixed to 2, i.e $\delta(v) = v^2$. Indeed, the Euclidean distance implies interesting simplifications of the speed. First, the speed becomes:

$$\tau_2(v) = \frac{1}{k(v) - 1}$$

where $k(v) = \frac{\overline{(v-\overline{v})^4}}{\overline{(v-\overline{v})^2}^2}$ is the Pearson kurtosis. Second, we can rewrite the speed directly as a function of of feature components q and s:

$$\tau_2(v) = \frac{1}{\frac{\overline{(q-s)^4}}{\overline{(q-s)^2}^2} - 1}$$

Since q and s have the same pdf e, we have $\overline{(q-s)^2} = 2\overline{q^2} - 2\overline{q}^2$, $\overline{(q-s)^4} = 2\overline{q^4} - 8\overline{q^3}\,\overline{q} + 6\overline{q^2}^2$, and thus:

$$\tau_2(q) = \frac{1}{\frac{\overline{q^4} - 4\overline{q^3}\,\overline{q} + 3\overline{q^2}^2}{2(\overline{q^2} - \overline{q}^2)^2} - 1}$$

We now introduce the α parameter of similarity measures $S_{\alpha,2}$. By adapting the derivation of Section 3 where $v = q - s$, to the case where $v = q^\alpha - s^\alpha$ (q and s are assumed positive), we deduce the following rectified precision:

$$r_2(q) = \frac{\sqrt{N_f}\,\overline{m_i^2}}{2\overline{q^2}} \frac{\alpha((2\alpha-1)\overline{q^{2\alpha-2}} + \alpha\overline{q^{\alpha-1}}^2 - (\alpha-1)\overline{q^{\alpha-2}}\,\overline{q^\alpha})\overline{q^2}}{2\sqrt{2}\sqrt{\overline{q^{4\alpha}} - 4\overline{q^{3\alpha}}\,\overline{q^\alpha} + 3\overline{q^{2\alpha}}^2 - 2(\overline{q^{2\alpha}} - \overline{q^\alpha}^2)^2}} \quad (12)$$

The rectified precision in (12) is written as the product of two terms, as in (8). Similarly, the new speed can be defined as:

$$\tau_\alpha(q) = \frac{\alpha^2((2\alpha-1)\overline{q^{2\alpha-2}} + \alpha\overline{q^{\alpha-1}}^2 - (\alpha-1)\overline{q^{\alpha-2}}\,\overline{q^\alpha})^2\overline{q^2}^2}{8(\overline{q^{4\alpha}} - 4\overline{q^{3\alpha}}\,\overline{q^\alpha} + 3\overline{q^{2\alpha}}^2 - 2(\overline{q^{2\alpha}} - \overline{q^\alpha}^2)^2)} \quad (13)$$

The speed τ_α is invariant with respect to the scale of q. The search for the best similarity measure within $S_{\alpha,2}$ is equivalent to maximizing the speed $\tau_\alpha(q)$ with respect to α.

7.1 Weibull

From Section 5, a more realistic pdf than exponential is the Weibull distribution:

$$w(q|s,a) = \frac{a}{s} q^{a-1} e^{-\frac{q^a}{s}} I_{q>0}$$

obtained by applying the power function $x^{\frac{1}{a}}$ where x is an exponential random variable. From [4], its mean is $s^{\frac{1}{a}}\Gamma(\frac{1}{a}+1)$, and its variance $s^{\frac{2}{a}}(\Gamma(\frac{2}{a}+1) - \Gamma^2(\frac{1}{a}+1))$. More generally, its moment of order p is:

$$\overline{q^p} = s^{\frac{p}{a}}\Gamma(\frac{p}{a}+1)$$

After substitution of the previous moment equation in the speed (13), we deduce the speed for a Weibull pdf as a function of α:

$$\tau_\alpha(Weibull) = \frac{(\Gamma(\frac{2}{a}+1) - \Gamma^2(\frac{1}{a}+1))^2}{8} \frac{\alpha^2 U_\alpha^2}{V_\alpha}$$

with $U_\alpha = (2\alpha-1)\Gamma(\frac{2\alpha-2}{a}+1) + \alpha\Gamma^2(\frac{\alpha-1}{a}+1) - (\alpha-1)\Gamma(\frac{\alpha-2}{a}+1)\Gamma(\frac{\alpha}{a}+1)$ and $V_\alpha = \Gamma(\frac{4\alpha}{a}+1) - 4\Gamma(\frac{3\alpha}{a}+1)\Gamma(\frac{\alpha}{a}+1) + 3\Gamma^2(\frac{2\alpha}{a}+1) - 2(\Gamma(\frac{2\alpha}{a}+1) - \Gamma^2(\frac{\alpha}{a}+1))^2$. The speed $\tau_\alpha(Weibull)$ is defined only when $a > 0$, $\alpha > -\frac{a}{4}$, $\alpha > -a+2$ and $\alpha > 1 - \frac{a}{2}$.

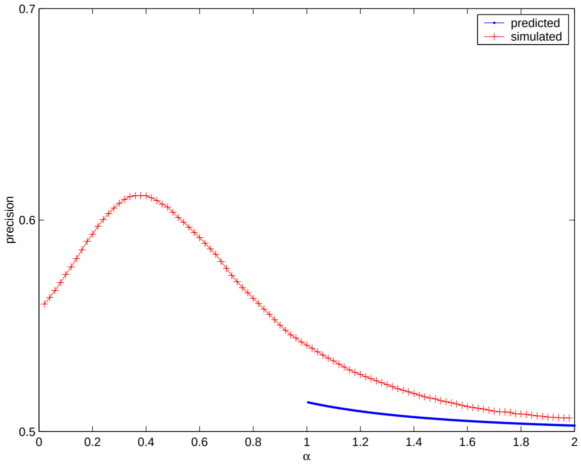

Figure 13: Comparison between predicted and observed average precision in the exponential case when α varies in $S_{\alpha,2}$.

As shown in α-diagrams of Fig. 13, where $a = 1$, the speed is defined only for $\alpha > 1$. Nevertheless as shown by simulation, it exists a value of α close to $\frac{1}{2}$ where the simulated precision is maximum. This can explain the results descibed in Sec. 2.

7.2 Design of Ground-truth Databases

In the previous sections, synthetic databases have been used to validate the model by showing the fit between the predicted and simulated average precisions. Of course there is not a good fit between simulation and theory for all kind of synthetic databases. The synthetic database must be designed in such a way that the measure of the simulated average precision is reliable.

From our experiments, we have inferred several specifications the synthetic database must follow:

- The standard deviation of the feature components should be 5 times higher than the difference of means between two feature components.
- The number of images N_i per class should be higher than 200.
- The size of the feature vector N_f should be higher than 100. If not the average precision becomes too noisy.

The last two rules are very important to help in designing real ground-truth databases as used in Section 2. If these specifications are followed, it is then possible to generalize, to other image databases similar in content to the ground truth-database, the best similarity measure obtained on the ground-truth database.

8. CONCLUSIONS

We have shown how to derive the best similarity measure in a set of similarity measures based on a proposed model of image retrieval systems. The statistical model assumes independent feature components with same parametric model for the pdfs but different means. To simplify the explanations, we have assumed that the image database has two classes, but the proposed model can be extended without difficulties to a larger number of classes. This model is partially tested on real data by fitting of feature components over a dedicated database. For fitting color histograms, family (7) which generalizes Gauss, Weibull and Gamma pdfs seems convenient.

We have illustrated this derivation on several examples assuming Gaussian, exponential, uniform, and Weibull pdfs for the feature variability. This derivation is useful when the feature variability can be modeled by a known family of parametric pdfs. To our knowledge, very little has been done on modeling the variability of features used in image retrieval systems. Nevertheless, this is probably a very important subject to study if we want to develop better image retrieval systems.

If there is no way to derive a model for the feature variability, a search of the best similarity measure can be performed on a ground-truth database. This is an optimization problem over the parameters of a set of similarity measures. The parameterized set of similarity measures (1) is very convenient for such a purpose. In particular, we have shown with these similarity measures interesting increases of average precision. We noticed that the best similarity measure seems mainly related to the feature space rather than to the choice of the database. From our tests on synthetic databases, we inferred several rules than must be followed in designing ground-truth databases for reliable conclusions on the choice of the best similarity measure. If these specifications are followed, it is then possible to generalize, to other image databases similar in content, the best similarity measure obtained on the ground-truth database. To tackle the over-fitting problem, rather than to use the central limit theorem as we did, an interesting possibility is to use the Berry-Eséen theorem [4] to obtain in which range the average precision is.

The first limit of the approach is that we assume independent feature vector components. The second limit is that the variances of the feature components are assumed equal. We are planing to relax these assumptions in future works. The main perspective of this work is to develop algorithms for adapting the similarity measure parameters by performing learning.

Acknowledgments

The authors are grateful to David B. Cooper from Brown University, USA, for valuable discussions and interesting comments.

9. REFERENCES

[1] S. Aksoy and R. M. Haralick. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, 22(5):563–582, 2001.

[2] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5), 1999.

[3] A. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, 1996.

[4] V. Koroliouk, N. Portenko, A. Skorokhod, and A. Tourbine. *Aide-mémoire de théorie des probabilités et de statistique mathématique.* Editions Mir, Moscou, 1983.

[5] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *IEEE International Conference on Computer Vision (ICCV'95)*, pages 786–793, Cambridge, USA, June 1995.

[6] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[7] N. Vasconcelos and A. Lippman. A bayesian framework for semantic content characterization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, USA, 1998.

[8] N. Vasconcelos and A. Lippman. A unifying view of image similarity. In *Proceedings of IEEE International Conference on Pattern Recognition (ICPR'00)*, Barcelona, Spain, 2000.

[9] C. Vertan and N. Boujemaa. Upgrading color distributions for image retrieval: Can we do better? In *Advances in Visual Information Systems. Proc.4'th Int'l Conf., VISUAL*, 2000.