Retrieval of Difficult Image Classes Using SVM-Based Relevance Feedback

Marin Ferecatu
INRIA Rocquencourt
Domaine de Voluceau,
Rocquencourt, B.P. 105
78153 Le Chesnay Cedex,
France
Marin.Ferecatu@inria.fr

Michel Crucianu
INRIA Rocquencourt
Domaine de Voluceau,
Rocquencourt, B.P. 105
78153 Le Chesnay Cedex,
France
Michel.Crucianu@inria.fr

Nozha Boujemaa
INRIA Rocquencourt
Domaine de Voluceau,
Rocquencourt, B.P. 105
78153 Le Chesnay Cedex,
France
Nozha.Boujemaa@inria.fr

ABSTRACT

User-defined classes in large generalist image databases are often composed of several groups of images and span very different scales in the space of low-level visual descriptors. The interactive retrieval of such image classes is then very difficult. To address this challenge, we propose and evaluate here two general improvements of SVM-based relevance feedback methods. First, to optimize the transfer of information between the user and the system, we focus on the criterion employed by the system for selecting the images presented to the user at every feedback round. We put forward a new active learning selection criterion that minimizes redundancy between the candidate images shown to the user. Second, for image classes having very different scales, we find that a high sensitivity of the SVM to the scale of the data brings about a low retrieval performance. We then argue that insensitivity to scale is desirable in this context and we show how to obtain it by the use of specific kernel functions. The experimental evaluation of both ranking and classification performance on several image databases confirms the effectiveness of our selection criterion and of the use of kernels that reduce the sensitivity of SVMs to the scale of the data.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback*

General Terms

Algorithms, Experimentation, Performance

Keywords

image retrieval, sample selection, active learning, reduction of redundancy, kernel function

MIR'04, October 15–16, 2004, New York, New York, USA.

1. INTRODUCTION

The cost of providing rich and reliable textual annotations for images in large databases, as well as the "linguistic gap" associated to these annotations, explains why the retrieval of images based on their **visual** content (content-based image retrieval, CBIR) is of high interest today [11].

Recently, the concept of "semantic gap" has been extensively used in the CBIR research community to express the discrepancy between the low-level features that can be readily extracted from the images and descriptions that are meaningful for the users of the search engine. The automatic association of such descriptions to the low-level features is currently only feasible for very restricted domains and applications. When searching more generic image databases, one solution for reducing the semantic gap is to cut a search session into several consecutive retrieval rounds (iterations) and let the user provide feedback regarding the results of every retrieval round (relevance feedback, RF).

The RF method embodied in a search engine should operate in real time and should maximize the ratio between the quality (or relevance) of the results and the amount of interaction between the user and the system. An RF method (see [20] for a review) is usually defined by two components, a learner and a selector. At every feedback round, the learner uses the images marked as "relevant" or "irrelevant" by the user to re-estimate the target of the user. Given the current estimation of the target, the selector chooses the images for which the user is asked to provide feedback during the next round.

The task of the learner is very difficult in the context of RF (see [5], [20]) since training examples are scarce (their number is usually lower than the number of dimensions of the description space), the training set is heavily imbalanced (there are often many more "irrelevant" examples than "relevant" ones) and both training and evaluation must be performed in real time. Much recent work is based on support vector machines (SVM, [18], [15]) because they avoid too restrictive assumptions regarding the data (e.g. that classes should have elliptic shape), are very flexible (can be tuned by kernel engineering) and allow fast learning and evaluation for medium-sized databases.

We would like to emphasize here the fact that RF was applied to **two important problems** of different nature. The **first** and most common type of problem consists in finding images in a specific target set; the focus is on ranking

most of the "relevant" images before the "irrelevant" ones rather than on finding a frontier between "relevant" and "irrelevant" images. The **second** use of RF is in defining a class of images for extending a textual annotation of some images in the class to the others; clearly, in this case the focus is on identifying a good frontier between the class of interest and the other images.

For each kind of problem a specific evaluation method should be used: in the first case we must measure the speed of improvement of the ranking (the precise ranking of the "relevant" or of the "irrelevant" images is usually unimportant), while in the second case we have to evaluate the speed of improvement of the classification.

In much of the work on RF, the images for which the user is asked to provide feedback at the next round were simply those that were currently considered by the learner as (potentially) the most relevant; also, in some cases these images are randomly selected. An interesting step ahead was the introduction in [17] and [16] of an **active learning** framework for RF using SVMs. We put forward here a significant improvement, consisting in a selection criterion that reduces the redundancy between the images for which the user is asked to provide feedback; our criterion encourages the selection of images that are far from each other in the space of low-level visual descriptors and thus allows for a better exploration of the current frontier.

The image classes found in generalist databases are the potential targets of the users of an RF system. These classes can have various, complex shapes and span very different scales (from very compact to very spread) in the space of low-level visual descriptors. If the learner strongly depends of a scale parameter, no value for this parameter will be adequate for all the classes in a database; also, with very few training examples, the scale parameter is difficult to tune online to suit the current class. Such sensitiveness to the scale of the data does occur for SVMs when some kernels (such as the RBF one) are employed. We argue that invariance (or at least low sensitivity) to scale is then an important desirable feature of the learner in an RF context and we propose to use specific kernels that let SVMs achieve this.

To prepare the presentation of experimental results in the last two sections of the paper, in the next section we explain our choice of ground-truth databases for evaluating RF methods and we briefly describe the low-level image features we employ.

Our active selection criterion with a reduction of the redundancy is put forward in section 3 and compared to other criteria. In section 4 we compare the results obtained with different kernels and we show why kernels that produce scale-invariance of the SVMs should be preferred. The comparisons presented in sections 3 and 4 concern both ranking (first type of problem mentioned earlier) and classification performance (second type of problem).

2. SETTING OF THE STUDY

2.1 Databases for the evaluation of RF

A specific RF method can be developed and evaluated on a particular application, with a well-defined scenario and a well-identified group of users. Knowing the specific assumptions concerning the application, the scenario and the users may help optimizing the RF method. It is nevertheless important to find improvements to RF methods that are relatively general and apply to many contexts. Evaluating such improvements by experimenting with users is very difficult to set up, since it would require the cooperation of many different groups of people in various contexts.

The common alternative is to use image databases for which a ground truth is available; this ground truth usually corresponds to the definition of a set of mutually exclusive image classes, covering the entire database. Of course, for a ground-truth database an user can often find many other classes that overlap those of the ground truth, so the evaluation of a retrieval method on such a database cannot be considered exhaustive even with respect to the content of that single database.

To cover a wide range of contexts, it is very important to use several ground-truth databases and to have characteristics that differ not only among these databases, but also among the classes of each database. Note that by finding correlations between the results of the RF methods and the characteristics of some classes or databases, one can identify ways for adapting RF to a specific context.

Relevance feedback methods must help reducing the semantic gap. It may then be important for evaluating RF to avoid having in the ground-truth databases too many "trivial" classes, i.e. for which simple low-level visual similarity is a sufficient classification criterion (this may be the case for classes produced for evaluating simple queries by example), because such classes may severely bias the results.

With these criteria in mind, we selected several ground-truth databases for the evaluation:

- GT72 is composed of the 52 most difficult classes from the Columbia color database, each class containing 72 different views of an object on an uniform background. There is enough visual variability within every class of this database and, at the same time, the identity of each class is not subject to interpretation. The classes are also sufficiently large to allow for a pertinent use with RF.

- GT100 has 9 classes, each composed of 100 images selected from the Corel database. While the high-level semantics of each class are clearly defined, there is a strong low-level visual diversity within each class. This makes the GT100 database difficult for a QBVE approach but a good candidate for search with RF.

- GT30 (70 classes, each having 30 images) and GT9 (246 classes, each with 9 images only) were built from several sources (Web Museum, Corel, Vistex). GT9 mainly contains many "simple" classes and was originally developed for the evaluation of image features, while GT30 is composed of both "simple" classes and more difficult ones, i.e. having more low-level visual diversity.

By studying these databases, we found as a significant common characteristic the fact that the spatial size of the various classes (in the space of low-level visual descriptors) covers an important range of different scales: 1 to 7 for GT72, 1 to 8 for GT100, 1 to 9 for GT30 and 1 to 4 for GT9. We evaluated the spatial size of a class by computing both the mean distance between elements and the mean distance between an element and the center of gravity of the class.

More generally, such significant changes in spatial scale can occur from one database to another, from one class to

another within the same database or even between parts of the frontier of a same class. For user-defined classes in a bigger database associated to a real retrieval scenario, one should expect even larger changes in spatial scale from one class to another. This implies that a low sensitivity to the scale of the data could be a desirable feature of the learner employed for RF; we shall see in section 4 that this is indeed the case.

2.2 Image content description

In this section we present the image descriptors we use and we stress the important connection that exists between the quality of the image descriptors and relevance feedback.

We employ weighted color histograms described in [19], [2] using the Laplacian and local probability as pixel weighting functions. Weighting functions bring additional information into the histograms (e.g. local shape or texture), which is an important principle in building reliable image signatures. The resulting integrated signatures generally perform better than a combination of classical, single-aspect features.

To describe the shape content of an image we use a histogram based on the Hough transform, which gives the global behavior along straight lines in different directions. Texture feature vectors are based on the Fourier transform, obtaining a distribution of the spectral power density along the frequency axes. This signature performs well on texture images and, used in conjunction with other image signatures, can significantly improve the overall behavior. The resulting joint feature vector has more than 600 dimensions.

2.3 Dimensionality reduction

The very high number of dimensions of the joint feature vector can make RF impractical even for medium-size databases. Also, the higher the dimensionality of the description space, the more difficult is the task of the learner. In order to reduce the dimension of the feature vectors, we use linear principal component analysis (PCA), which is actually applied separately to each of the image features previously described.

We evaluate the retrieval performance of the resulting image descriptors in a QBVE context by building a precision-recall diagram for each database. After a reduction in dimension of about 5 times, we remain within a 5% overall loss of quality in the precision-recall diagrams.

We expected kernel PCA ([15]) to better focus on relevant nonlinear "dimensions"; this should indeed be the case when the manifold spanned by the images is very low-dimensional but significantly nonlinear. However, when comparing KPCA to linear PCA we noticed that the first did not perform so well on the generalist image databases we are using, suggesting that the previous assumption is wrong in these cases.

If all the classes were known a priori, then other methods such as discriminant analysis would be more appropriate for reducing the dimension of the description space. Such an assumption cannot be made in real situations where the classes are defined interactively by the users, so we also avoided making it here, for the ground-truth databases we used in our evaluation.

3. REDUCTION OF THE REDUNDANCY

In order to maximize the ratio between the quality (or relevance) of the results and the amount of interaction between the user and the system, the selection of images for which the user is asked to provide feedback at the next round must be carefully studied.

Interesting ideas were introduced in [9] and [8], where the problem under focus is the iterative search for one specific image in a database (*target* search); at every round, the user is required to choose, between the two images presented by the engine, the one that is closest to the target image. The selection criterion put forward in this case attempts to identify at every round the most informative binary selections, i.e. those that are expected to maximize the transfer of information between the user and the engine (or remove a maximal amount of uncertainty regarding the target). We consider that this criterion translates into two complementary conditions for the images in the selection: each image must be ambiguous given the current estimation of the target and the redundancy between the different images has to be low.

Unfortunately, the entropy criterion employed in [9], [8] does not scale well to the search of images in a larger set (*category* search) and to the selection of more than 2 images. Computational optimizations must be found, relying on the use of specific learners and, possibly, specific search contexts.

Based on the definition of active learning (see for example [7]), the selection of examples for training SVMs to perform general classification tasks is studied in [4]. When the classification error increases with the distance between the misclassified examples and the frontier (a "soft margin" is used for the SVM), the authors interestingly distinguish two cases: early and late stages of learning.

In the early stages, the classification of new examples is likely to be wrong, so the fastest reduction in generalization error can be achieved by selecting the example that is farthest from the current estimation of the frontier. During late stages of learning, the classification of new examples is likely to be right but the margin may be suboptimal, so the fastest reduction in error can be achieved by selecting the example that is closest to the current estimation of the frontier. Note that, according to the classical formulation of active learning, the authors only consider the selection of single examples for labeling (for addition to the training set) at every round.

For SVM learners, several selection criteria are presented in [17] and applied to content-based text retrieval with relevance feedback. The simplest (and computationally cheapest) of these criteria consists in selecting the texts whose representations (in the feature space induced by the kernel) are closest to the hyperplane currently defined by the SVM. We shall call this simple criterion the selection of the "most ambiguous" (MA) candidate(s).

This selection criterion is justified in [17] by the fact that knowledge of the label of such a candidate halves the version-space. In this case, the version space is the set of parameters of the hyperplanes in feature space that are compatible with the already labeled examples. The proof of this result assumes that the version space is not empty and that, in the feature space associated to the kernel, all the images of vectors in the input space have constant norm.

These assumptions will hold with appropriate choices for the kernel and for the bound (C) on the parameters of the SVM (the α). In order to minimize the number of learning rounds, the user is asked to label several examples at every round and all these examples are selected according to the MA criterion. In [16] the MA selection criterion is applied

to CBIR with relevance feedback and shown to produce a faster identification of the target images than the selection of random images for labeling.

While the MA criterion provides a computationally effective solution to the selection of the most ambiguous examples, when used for the selection of more than one candidate example it does not remove the redundancies between the candidates.

Suppose now that x_i and x_j are the input space representations of two candidate examples and consider kernels K such that $K(x_i, x_j)$ monotonously decreases with an increase of the distance $d(x_i, x_j)$ (this is the case for most common nonlinear kernels). For such kernels, we propose here to require a low value for $K(x_i, x_j)$ as an **additional condition of low redundancy**. This obviously encourages the selection of unlabeled examples that are far from each other in input space, allowing us to better explore the current frontier between "relevant" and "irrelevant" images.

If the kernel K is further inducing a Hilbert structure on the feature space and if all the $\phi(x)$ (the images in feature space of the vectors x in input space) have constant norm, then the additional condition stated above corresponds to a requirement of (quasi-)orthogonality between $\phi(x_i)$ and $\phi(x_j)$, since $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. In this case, our condition of low redundancy can also be justified by reference to the version space account suggested in [17]: redundancy is minimized when the hyperplanes associated to the individual examples are orthogonal and are thus complementary to each other in halving the version space. This is why we shall call this criterion the selection of the "most ambiguous and orthogonal" (MAO) candidates even if, as we shall see in section 4, this name is not appropriate for certain kernels.

The low redundancy condition stated above can be modified for kernels that don't satisfy our hypotheses (that is, $K(x_i, x_j)$ monotonously decreases with an increase of $d(x_i, x_j)$ and, if defined, $\|\phi(x)\| = \text{constant}, \forall x$). As an example, for positive definite kernels that don't satisfy the condition $\|\phi(x)\| = \text{constant}, \forall x$, one should require a low value for $K(x_i, x_j)/\sqrt{K(x_i, x_i)\ K(x_j, x_j)}$ rather than for $K(x_i, x_j)$. In a standard classification context, a similar "diversity" condition for the selected examples was put forward in [3] and evaluated on several benchmark classification problems from the UCI database.

To implement our MAO criterion, we first perform an MA selection of a larger set of unlabeled examples. Then, for kernels who satisfy our hypotheses (see above), we build the MAO selection by iteratively choosing as a new example the vector x_j that minimizes the highest of the values taken by $K(x_i, x_j)$ for all the x_i examples already included in the current MAO selection. This can be expressed as

$$x_j \ = \ \text{argmin}_{x \in S} \ \max_i \ K(x, x_i) \qquad (1)$$

where S is the set of images not yet included in the current MAO selection and x_i, $i = 1, \ldots, n$ are the already chosen candidates.

The size of the set of unlabeled examples pre-selected with MA should be proportional to the number ws (window size) of images for which the user is asked to provide feedback at the next round. The low redundancy condition is simply ignored if this size is too small and the ambiguousness condition is ignored if it is too large. In our experiments we found that a value of $2 \cdot ws$ was a good compromise for the different databases and classes, so we used it for obtaining the results presented in the following.

It is easy to see that the MAO criterion can be extended to reduce redundancies between the examples selected during subsequent RF rounds, by regarding all these examples as already chosen candidates when computing (1). This additional constraint may be important in situations where the number of labeled examples is much lower than the dimension of the input space and the classes are restricted in most directions. However, in the experiments presented below we did not attempt to reduce redundancies between subsequent rounds of RF.

We note that the MA criterion in [17], [16] is the same as the one put forward in [4] for the late stages of learning. This clarifies the fact that the MA criterion relies on two important further assumptions: **first**, the prior on the version space is rather uniform; **second**, the solution found by the SVM is close to the center of gravity of the version space. The second assumption can be relieved by using Bayes Point Machines [12] instead of SVMs or one of the more sophisticated criteria put forward in [17], albeit at a higher computational cost.

However, in the early stages of an RF session the frontier will usually be very unreliable and, depending on the initialization of the search and the characteristics of the classes, may be much larger than the target class (there are much fewer examples than dimensions in the description space). It follows that the first assumption may not hold in the early stages of learning. In such cases, selecting those unlabeled examples that are currently considered by the learner as (potentially) the most relevant can sometimes produce a faster convergence during the first few rounds of RF. For this reason, we added to our comparisons the following criteria: select the "most positive" unlabeled examples according to the current decision function of the SVM (denoted as MP) and select the "most positive and orthogonal" unlabeled examples (denoted as MPO). The MPO criterion adds to MP the condition of low redundancy previously described.

When comparing the MP criterion to the suggestion in [4] for the early stages of learning, we see that we only focus on the examples for which the values taken by the decision function of the SVM are maximal and completely ignore the examples for which these values are minimal; this is because of the asymmetry of the retrieval context: in general, the number of relevant items is expected to be much lower than the number of irrelevant items.

We performed comparisons between the four selection criteria on the ground-truth databases we retained. For the GT72, GT100 and GT30 databases, at every feedback round the (emulated) user must label as "relevant" or "irrelevant" all the images in a window of size $ws = 9$. The window size is reduced to 4 for the GT9 database.

A search session is initialized by considering one "relevant" example and $ws - 1$ "irrelevant" examples. Every image in every class serves as the initial "relevant" example for a different RF session, while the associated initial $ws - 1$ "irrelevant" examples are randomly selected. For reasons that will become apparent in section 4, we use for the comparisons presented here the triangular kernel (also introduced in section 4).

We began by evaluating the different selection criteria on the **first type of problem** mentioned in the introduction: finding items in a specific target set, by focusing on ranking

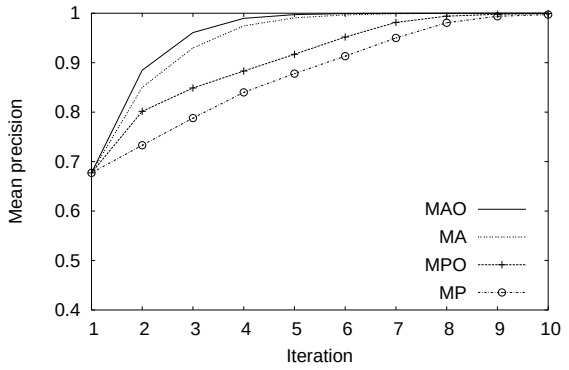

Figure 1: Evolution of the mean precision obtained with the different selection criteria on the GT72 database.

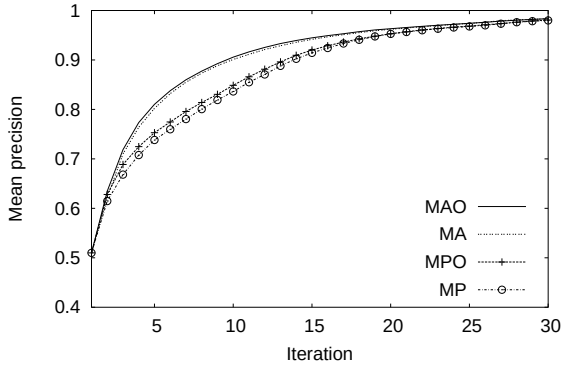

Figure 2: Evolution of the mean precision obtained with the different selection criteria on the GT100 database.

most of the "relevant" images before the "irrelevant" ones rather than on finding a frontier between the class of interest and the other images. Since the only information available concerns class membership (crisp value), we do not consider important here the precise ranking of the "relevant" or of the "irrelevant" images.

In order to evaluate the speed of improvement of this ranking, we must use a measure that does not give a prior advantage to one selection criterion. For example, by taking into account already labeled images plus those selected for being labeled during the current round, we should obviously favor the MP and MPO criteria over MA and MAO. We decided to use instead the following precision measure: at every RF round, we count the number of truly "relevant" images found in the N images considered as most positive by the current decision function of the SVM (N being the number of images in each class, fixed for each of the ground-truth databases we studied).

The evolution of the mean precision during successive RF iterations (rounds) on the GT72 and GT100 databases are presented in Fig. 1 and 2. The "mean precision" value shown is obtained as the mean value over all feedback rounds (each image in the database is used to initiate a new feedback round as described above).

Clearly, the reduction of the redundancy between the images selected for labeling improves the results, both for MAO

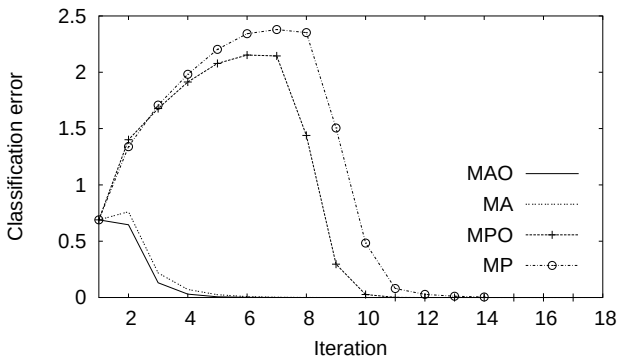

Figure 3: Evolution of the classification error obtained with the different selection criteria on the GT72 database.

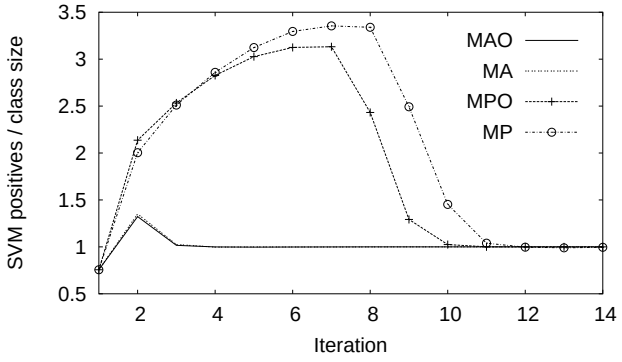

Figure 4: Evolution of the ratio between the images considered by the SVM as "relevant" and the number of images in a class for the different selection criteria on the GT72 database.

with respect to MA and for MPO with respect to MP. Also, in these cases the MA and MAO selection criteria compare favorably to the MP and MPO criteria.

The **second type of problem** mentioned in the introduction consists in finding a frontier between "relevant" and "irrelevant" images, which can be important for extending a textual annotation of some images in the "relevant" class to the others. In this case, we have to evaluate the speed of improvement of the classification.

The classification error is defined here as $n/N + (N - p)/N$, where N is the class size, n is the number of false positives and $N - p$ is the number of false negatives. In Fig. 3 we can see the evolution of the classification error for the different selection criteria on the GT72 database. As expected, the convergence is fastest for the MAO selection criterion, followed by the MA criterion.

To better understand the behavior of the different selection criteria, we also studied the evolution of the ratio between the images considered by the SVM as "relevant" and the number of images in the class (this ratio should be 1 when SVM has found the entire class).

The results obtained on the GT72 database are shown in Fig. 4. The convergence is significantly faster for the MAO and MA criteria, which can be explained by the fact that MP and MPO do not focus on the frontier.

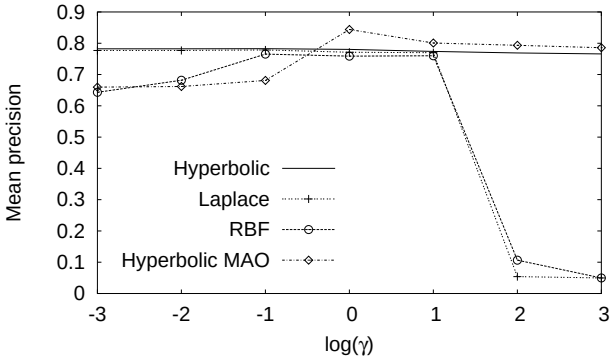

Figure 5: Sensitivity of the SVM to a scale parameter for the different kernels on the GT72 database.

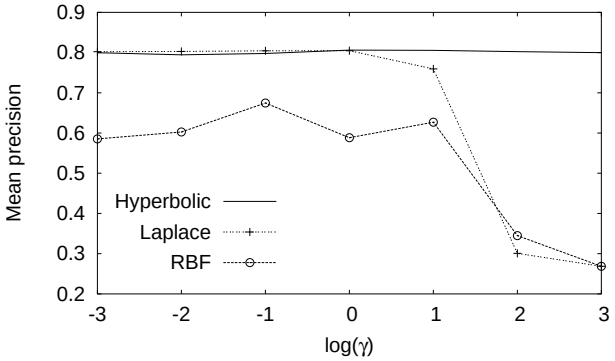

Figure 6: Sensitivity of the SVM to a scale parameter for the different kernels on the GT100 database.

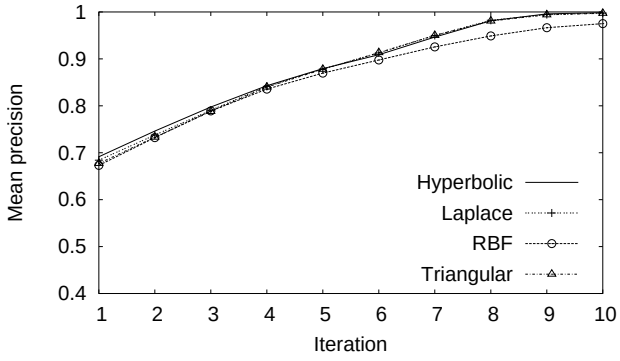

Figure 7: Comparison of the different kernels on the GT72 database, with the MP selection criterion.

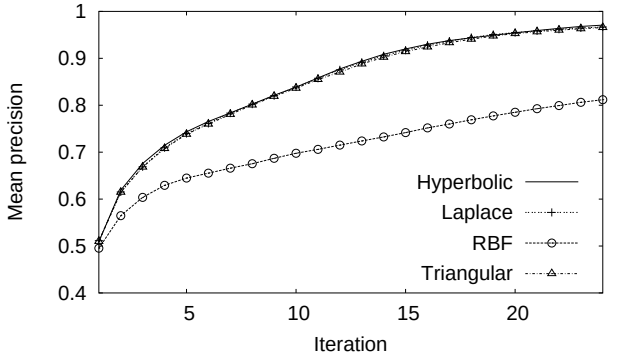

Figure 8: Comparison of the different kernels on the GT100 database, with the MP selection criterion.

The results we obtained on all the other databases we considered are similar to those in Fig. 3 and 4. These results suggest that the MAO selection criterion should be strongly preferred whenever RF is used for finding a reliable frontier between "relevant" and "irrelevant" images (or interactively learning new "visual concepts").

4. INVARIANCE TO SCALE

During the study of several ground-truth databases we found that the spatial size of the various classes (in the space of low-level visual descriptors) often covers an important range of different scales (1 to 7 for the GT72 database, 1 to 8 for the GT100 database). We expect yet more significant changes in spatial scale to occur from one database to another, from one user-defined class to another within a large real-world image database and even between parts of the frontier of some classes. A too strong sensitivity of the learner to the scale of the data could then strongly limit its applicability in an RF context.

For SVM classifiers, sensitivity to scale has two sources: the scale parameter of the kernel and the C bound on the α coefficients. We focus here on the first source of sensitivity, the second one being usually less constraining (the C bound can be set in our retrieval context to some high value without significantly affecting performance).

The first kernel we consider is the Gaussian (or Radial Basis Function) one, $K(x_i, x_j) = \exp\big(-\gamma\|x_i - x_j\|^2\big)$. This classical nonlinear kernel, often employed by default, is highly sensitive to the scale parameter γ (the inverse of the variance of the Gaussian).

The use of the Laplace (or exponential) kernel, $K(x_i, x_j) = \exp\big(-\gamma\|x_i - x_j\|\big)$, was advocated in [6] for histogram-based image descriptors. In [13], this kernel was found to work better than the Gaussian kernel for CBIR with RF.

The hyperbolic kernel, $K(x_i, x_j) = 1/\big(\varepsilon + \gamma\|x_i - x_j\|\big)$, can be computed fast and we have already used it for RF with good results. The scale parameter is γ again; ε translates into a multiplicative constant plus a change in γ and is only used to avoid numerical problems (we set it to 0.001).

All the three kernels we mentioned are positive definite kernels and satisfy the condition $\|\phi(x)\| = \text{constant}$, $\forall x$. The triangular kernel, $K(x_i, x_j) = -\|x_i - x_j\|$, was introduced in [1] as a *conditionally* positive definite kernel, but the convergence of SVMs remains guaranteed with this kernel [14]. In [10] the triangular kernel was shown to have a very interesting property: it makes the frontier found by SVMs invariant to the scale of the data (within the limits set by the value of the C bound, but even these limits are less strong for the triangular kernel than for the Gaussian kernel).

Since the triangular kernel is not positive definite but only conditionally positive definite, the account provided in [17], [3] for the MA selection criterion does not hold for this kernel. For the same reason, the triangular kernel does not induce a Hilbert structure on the feature space, so in this case one should not speak of the "orthogonality" of vectors

in the feature space. However, since the value of $K(x_i, x_j)$ decreases with an increase of the distance $d(x_i, x_j)$, our explanation for the MAO criterion holds, as well as the justification of the MA criterion in [4]; we also continue to (abusively) use MAO as the name of the selection criterion.

For all the kernels we used the L1 norm because experimentally we found it to provide better results than L2. A few other dissimilarity measures (some of which don't have the properties of a metric) were used in the literature instead of $\|x_i - x_j\|$, mainly with the Gaussian kernel and sometimes for variable-length representations of the images. Some of these measures don't guarantee the convergence of the SVM and we preferred not to use them here.

The sensitivity of these kernels to the γ parameter with the MP criterion is shown in Fig. 5 for the GT72 database and in Fig. 6 for the GT100 database (the base of the logarithm is 10). In these two figures, the "mean precision" value shown is obtained as the mean over all the images in the database **and** over the first 15 feedback iterations. Also, comparisons between these kernels using the MP selection criterion are shown in Fig. 7 and 8; for the Gaussian, Laplace and hyperbolic kernels, the scale parameters were set to their optimal values for the database.

We only show comparisons employing the MP criterion because it is prevailing so far. Sensitivity to scale with the Gaussian, Laplace and hyperbolic kernels increases when the MAO criterion is used (see Fig. 5 for the effect of MAO on the hyperbolic kernel).

From Fig. 5 and 6 one can see that the Gaussian kernel is the one who produces the highest sensitivity to scale for the SVM. Since the classes present in a database often have significantly different spatial scales, any value for the scale parameter will be inadequate for many classes, so the results obtained with this kernel cannot be very good.

Comparatively, the use of the Laplace kernel reduces the sensitivity of the SVM to scale. With the Laplace kernel and the MP selection criterion, an increase of γ beyond 1 has a strong negative impact on the results, while a reduction of γ does not have significant consequences. This is explained by the fact that for small γ the Laplace kernel becomes similar to the triangular kernel.

With the MP selection criterion, the hyperbolic kernel produces a scale-invariance of the SVM within a large range of values for γ. However, as shown by the "Hyperbolic MAO" line in Fig. 5, this invariance is lost to some extent when the MAO selection criterion is employed.

We also compared the classification performance obtained with the different kernels. As shown in Fig. 9, with the triangular kernel the classification error converges to zero relatively fast (even with the MP criterion), while with the RBF kernel it continues to increase even after 20 feedback rounds. We didn't display the lines corresponding to the hyperbolic and Laplace kernels in Fig. 9 because convergence is almost as fast with these kernels as with the triangular kernel. Similar behaviors were found during the experiments on the other databases.

For real-world applications, the spatial scales of the user-defined classes cannot be known *a priori* and the scale parameter of a kernel cannot be easily adjusted online, so important variations between classes can be expected for the performance of RF-based retrieval if kernels such as the RBF one are employed. The scale-invariance obtained by the use of the triangular kernel (or, to a lesser extent, of the hyperbolic kernel) becomes then a highly desirable feature and, according to the experimental evidence we presented here, makes this kernel a very good alternative.

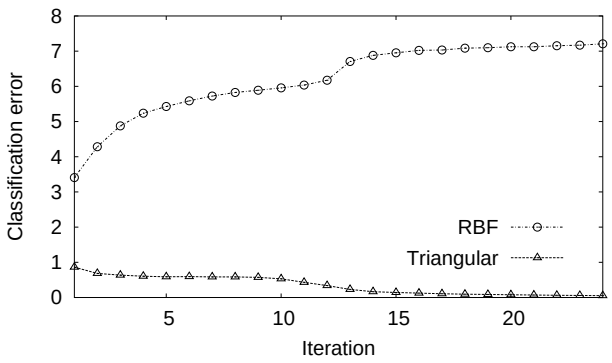

Figure 9: Comparison of the classification error for the RBF and the triangular kernels on the GT100 database.

5. CONCLUSION

In content-based image retrieval with relevance feedback, the criterion employed by the search engine for selecting the images presented to the user at every feedback round is very important for the transfer of information between the user and the system. Using SVMs as learners, we put forward an improved active learning selection criterion, based on a reduction of the redundancy between the images selected at every feedback round. By comparing this criterion to alternative criteria on several ground-truth databases, we have shown that it performs better in ranking most of the "relevant" images before the others.

We also took a fresh look at active learning for relevance feedback as a tool for learning complex concepts (in order to extend textual annotations or for other uses) and we presented experimental evidence that it speeds up both the convergence of the classification error to zero and the convergence of the frontier around the class of interest.

By studying several ground-truth databases, we found in the space of low-level visual descriptors significant changes in spatial scale between the various classes. Yet greater changes in scale are expected to occur for user-defined classes in real-world applications. We have shown that a high sensitivity of the learner to changes in the scale of the data strongly degrades its performance and limits its applicability in a relevance feedback context. For SVMs, we proposed to use specific kernel functions, such as the triangular kernel, that allow to obtain insensitivity to changes in scale and, as experiments on different databases show, keep performance at a very good level.

As an example of retrieval, in Fig. 10 we present the third screen of results returned by our system IKONA [2], using the methods presented in this paper, from a generalist database of 3670 images. In this case the user is looking for a rather difficult class: portraits that are paintings at the same time. With 4 positive and 10 negative examples, the images returned by the system are all portraits, but nevertheless they have very different characteristics: different backgrounds, different cloth colors, different overall texture.

Figure 10: Searching for portraits in a generalist database.

6. ACKNOWLEDGMENTS

Marin Ferecatu is partly financed by a grant of the Région Ile-de-France.

7. REFERENCES

[1] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, 1984.

[2] N. Boujemaa, J. Fauqueur, M. Ferecatu, F. Fleuret, V. Gouet, B. L. Saux, and H. Sahbi. Ikona: Interactive generic and specific image retrieval. In *Proceedings of the International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001)*, 2001.

[3] K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of ICML-04, International Conference on Machine Learning*, pages 59–66, August 2003.

[4] C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 111–118. Morgan Kaufmann, 2000.

[5] E. Y. Chang, B. Li, G. Wu, and K. Goh. Statistical learning for effective visual image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'03)*, pages 609–612, September 2003.

[6] O. Chapelle, P. Haffner, and V. N. Vapnik. Support-vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.

[7] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[8] I. J. Cox, M. L. Miller, T. P. Minka, T. Papathomas, and P. N. Yianilos. The Bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.

[9] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. An optimized interaction strategy for Bayesian relevance feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 553–558. IEEE Computer Society, 1998.

[10] F. Fleuret and H. Sahbi. Scale-invariance of support vector machines based on the triangular kernel. In *3rd International Workshop on Statistical and Computational Theories of Vision*, October 2003.

[11] T. Gevers and A. W. M. Smeulders. Content-based image retrieval: An overview. In G. Medioni and S. B. Kang, editors, *Emerging Topics in Computer Vision*. Prentice Hall, 2004.

[12] R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, 2001.

[13] F. Jing, M. Li, L. Zhang, H.-J. Zhang, and B. Zhang. Learning in region-based image retrieval. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2003.

[14] B. Schölkopf. The kernel trick for distances. In *Advances in Neural Information Processing Systems*, volume 12, pages 301–307. MIT Press, 2000.

[15] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[16] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM Press, 2001.

[17] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 999–1006. Morgan Kaufmann, 2000.

[18] V. Vapnik. *Estimation of dependencies based on empirical data*. Springer Verlag, 1982.

[19] C. Vertan and N. Boujemaa. Upgrading color distributions for image retrieval: can we do better? In *International Conference on Visual Information Systems (Visual2000)*, November 2000.

[20] X. S. Zhou and T. S. Huang. Relevance feedback for image retrieval: a comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.