Reducing the Redundancy in the Selection of Samples for SVM-based Relevance Feedback

Marin Ferecatu, Michel Crucianu, Nozha Boujemaa

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Reducing the Redundancy in the Selection of Samples for SVM-based Relevance Feedback

Marin Ferecatu — Michel Crucianu — Nozha Boujemaa

N° 5258

July 5, 2004

THÈME 3

ISSN 0249-6399

Reducing the Redundancy in the Selection of Samples for SVM-based Relevance Feedback

Marin Ferecatu , Michel Crucianu , Nozha Boujemaa

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Imedia

Rapport de recherche n° 5258 — July 5, 2004 — 20 pages

Abstract: In image retrieval with relevance feedback, the strategy employed by the system for selecting the images presented to the user at every feedback round has a strong effect on the transfer of information between the user and the system. Using SVMs, we put forward a new active learning selection strategy that minimizes redundancy between the images presented to the user and takes into account assumptions that are specific to the retrieval setting. Experiments on several image databases confirm the attractiveness of this selection strategy. We also find that insensitivity to the scale of the data is a desirable property for the SVMs employed as learners in relevance feedback and we show how to obtain such insensitivity by the use of specific kernel functions.

Key-words: image retrieval, sample selection, active learning, reduction of redundancy, kernel function

Réduction de la redondance dans la sélection des images pour le retour de pertinence avec SVM

Résumé : Dans la recherche d'images avec retour de pertinence, la stratégie employée par le système pour sélectionner les images présentées à l'utilisateur à chaque itération a une grande importance pour le transfert d'information entre l'utilisateur et le système. Utilisant les SVM, nous proposons une nouvelle méthode de sélection basée sur l'apprentissage actif, qui minimise la redondance entre les images que l'utilisateur doit marquer et tient compte d'hypothèses spécifiques au contexte de la recherche. L'évaluation sur plusieurs bases d'images confirment l'attractivité de la stratégie de sélection proposée. Nous avons également constaté que l'invariance du résultat de l'apprentissage à l'échelle des données est une propriété utile dans ce contexte et nous montrons comment obtenir une telle invariance par l'emploi de fonctions-noyaux spécifiques.

Mots-clés : recherche d'images par le contenu, sélection d'exemples, apprentissage actif, réduction de la redondance, fonction noyau

Contents

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].

In the early years of research in CBIR, the focus was on "query by visual example" (QBVE): a search session begins by presenting an example image (or sketch) to the search engine as a "visual query", then the engine returns images that are visually similar to the query image. 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, e.g. by qualifying images returned as either "relevant" or "irrelevant" (relevance feedback, RF). From this feedback, the engine progressively learns the visual features of the images the user is looking for in the current search session (the "target" of the user).

It must be noted here that the semantic gap should not be seen as the unique explanation for the difficulties encountered in retrieval by visual content: the "numerical gap" or the use of incomplete or confusing descriptions of the visual content, is a complementary cause of problems in retrieval. To reduce this numerical gap, one has to develop image signatures

(low-level image features) that are both rich and faithful. To make retrieval more efficient, these signatures should also be compact.

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 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 an estimation of the target, the selector chooses the images for which the user is asked to provide feedback at the next round.

The task of the learner is very difficult in the context of RF (see [5], [20]) because 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 argue here that invariance to the scale of the data is an important desirable feature of the learner in the RF context and we show that some kernels are better than others in achieving this.

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 important step ahead was the introduction in [17] and [16] of an *active learning* framework for RF using SVMs. We put forward here an improvement of this active learning strategy, based on a reduction of the redundancy between the images selected, and attempt to identify the "domain of preference" of different selection strategies.

We must emphasize that RF was applied to two 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. An interesting but less common use of RF is in defining a class of images and extending textual annotations 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 the next section we explain our choice of groundtruth databases for evaluating RF algorithms and we describe the image signatures we employ. The active selection strategy with a reduction of the redundancy is presented in Section 3 and compared to other strategies. In Section 4 we compare the results obtained with different kernels and we show why kernels that produce scale-invariance of the SVM should be preferred.

2 Setting of the study

2.1 Databases for RF methods evaluation

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 helps 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 users 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 groundtruth 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 groundtruth 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 groundtruth 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 groundtruth 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 "trivial" classes and was originally developed for the evaluation of image signatures in a QBVE context, while GT30 is composed of both "trivial" 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 size of the various classes covers an important range of different scales: 1 to 5 for GT72, 1 to 8 for GT100, 1 to 9 for GT30 and 1 to 4 for GT9. More generally, such changes in 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 scale from one class to another. It follows 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 (signatures) we use and we stress the important connection that exists between the quality of the image descriptors and relevance feedback.

Finding good image descriptors that can accurately describe the visual aspect of many different classes of images is a challenging task. Such descriptors are easier to compute for specialized databases, where specific prior knowledge can be used to devise a more dedicated description of the image content. On one side, there is the rather subjective problem of the visual content and, on the other side, there is the very practical need to find a good technical/mathematical description of this same visual content. Since there is no perfect description of visual content (even humans disagree when interpreting images), most methods try to find a good compromise in balancing the different aspects of image content.

Integrated color/texture histograms. Classical color histograms give a statistical description of the color content in the image, but without keeping any spatial information: all pixels are equally important, independently of their position. However, pixels having the same color are not similar if we consider their neighborhood in the image [19], [2]. Follows the idea of weighting each color by a measure giving its importance in the local context:

$$h(\mathbf{c}) = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} w(i,j)\delta(f(i,j) - \mathbf{c})$$

where $h(\mathbf{c})$ is the histogram value for the color $\mathbf{c}$, M and N are the dimensions of the image (in pixels), $f(i,j)$ is the (image) color for the pixel (i,j), $w(i,j)$ is the weighting function and $\delta(\cdot)$ is the Dirac distribution. By using different weighting functions and various color spaces one can obtain a family of image descriptors. There is a dual aspect in the definition of the weighting functions: emphasizing uniform areas or emphasizing non-uniform areas — which aspect is more important depends on the query.

Using as weighting function $w(i,j) = \Delta^2(i,j)$, where Δ is the Laplacian computed at the current pixel (i,j), emphasizes corners and edges in the image. We obtain in this case a *Laplacian weighted histogram*. Another weighting function is the probability of the current pixel color within a local window (*probability weighted histogram*). If this probability is small

then the pixel belongs to a zone in the image where the color activity is important, otherwise we have a rather uniform area. This function measures the local color dominance.

Weighting functions bring new 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.

Texture/shape descriptors. 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 (*Hough histogram*).

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 image features we use here are: the Laplacian weighted histogram, the probability weighted histogram, the Hough histogram, the Fourier histogram and a classical color histogram obtained in HSV color space. The joint feature vector has more than 600 dimensions.

We should stress here the fact that good signatures, which combine several aspects of the image content (color, texture and shape) in a single descriptor, are the base for more sophisticated search methods such as relevance feedback. RF should not be used to compensate for low quality image signatures, but to perform searches that by other principles would be more difficult if not impossible. RF is in fact very related to subjective image classification, and a key element are the signatures that contain much information while having a low dimension.

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. Two such diagrams are presented in Fig. 1 and 2. After a reduction in dimension of about 5 times, we remain within a 5% overall loss of quality in the precision-recall diagrams.

In preliminary experiments, we applied PCA to the vectors containing all the image features but, with a similar reduction in dimension, the overall retrieval results were not as good. We also 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 performed rather poorly for the generalist image databases we are working on, suggesting that the previous assumption is wrong in these cases.

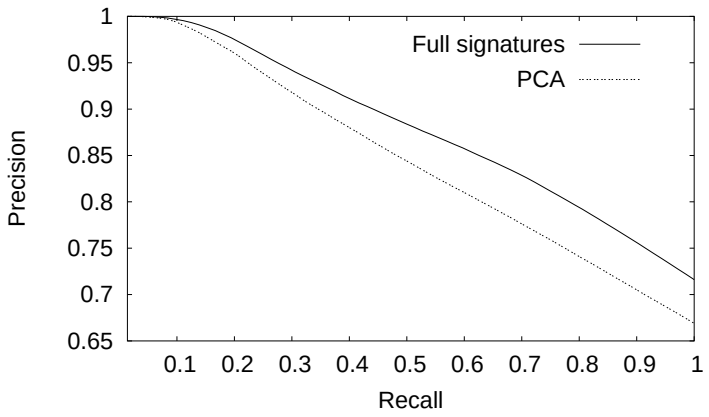

Figure 1: Result of a reduction in dimension with linear PCA on the GT72 database.

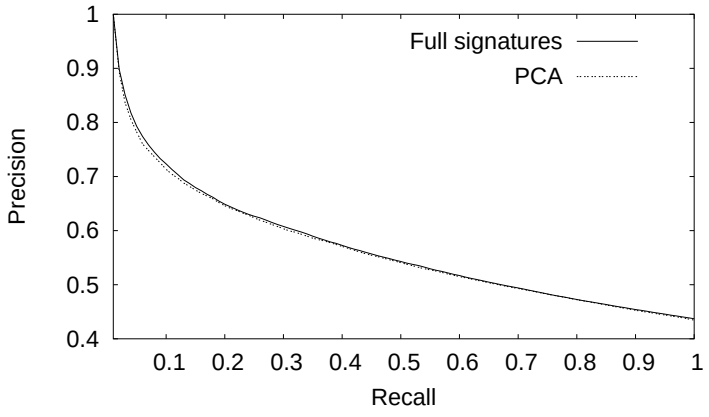

Figure 2: Result of a reduction in dimension with linear PCA on the GT100 database.

Note that 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 groundtruth databases we used in our evaluation.

3 Sample selection with redundancy reduction

3.1 Sample selection and active learning

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 strategy 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 entropic 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.

Also 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.

3.2 Reduction of the redundancy for active learning with SVM

While the MA criterion provides a computationally effective solution to the selection of the most ambiguous images (satisfying the first condition mentioned above), when used for the selection of more than one candidate image it does not remove the redundancies between the candidates (it does not satisfy the second condition).

We suggest here to translate this condition of low redundancy into the following additional condition: if x_i and x_j are the input space representations of two candidate images, then we require a low value for $K(x_i, x_j)$ (i.e. of the value taken by the kernel for this pair of images). If the kernel K is inducing a Hilbert structure on the feature space, if $\phi(x_i)$, $\phi(x_j)$ are the images of x_i, x_j in this feature space and if all the images of vectors in the input space have constant norm, then this additional condition corresponds to a requirement of (quasi-)orthogonality between $\phi(x_i)$ and $\phi(x_j)$ (since $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$). We shall call this criterion the selection of the "most ambiguous and orthogonal" (MAO) candidates (we shall see later that this name is not totally appropriate for certain kernels).

We should note that the MAO criterion can be extended to reduce redundancies between the examples selected during subsequent RF rounds. 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.

The MAO criterion has a simple intuitive explanation for kernels $K(x_i, x_j)$ that decrease with an increase of the distance $d(x_i, x_j)$ (which is the case for most common kernels): it encourages the selection of unlabeled examples that are far from each other in input space, allowing to better explore the current frontier. To implement this criterion, we first perform an MA selection of a larger set of unlabeled examples. Then, we build the MAO selection by iteratively choosing as a new example the x_j that minimizes the highest of the values taken by $K(x_i, x_j)$ for all the x_i examples that are already included in the current MAO selection.

In a general 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. The condition is justified by reference to the version space account suggested

in [17]: diversity is maximized when the hyperplanes associated to the individual examples are orthogonal and are thus complementary to each other in halving the version space.

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 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 of the frontier 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 criterion) 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.

3.3 Experimental evaluation of redundancy reduction

We performed comparisons between the four selection criteria on the groundtruth 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.

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 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

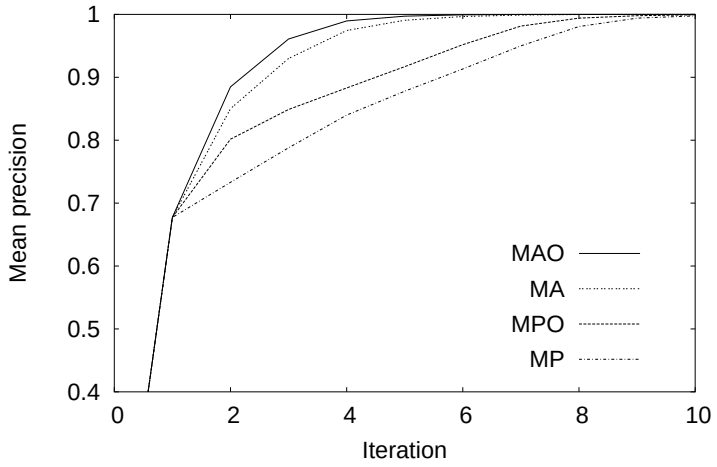

Figure 3: Comparison of the selection strategies on the GT72 database.

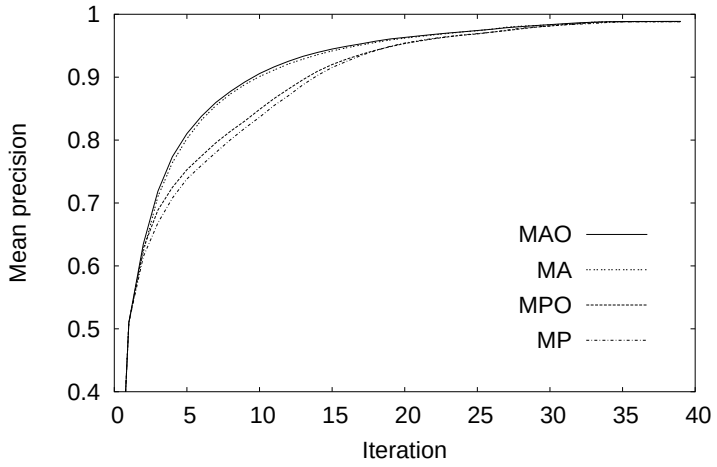

Figure 4: Comparison of the selection strategies on the GT100 database.

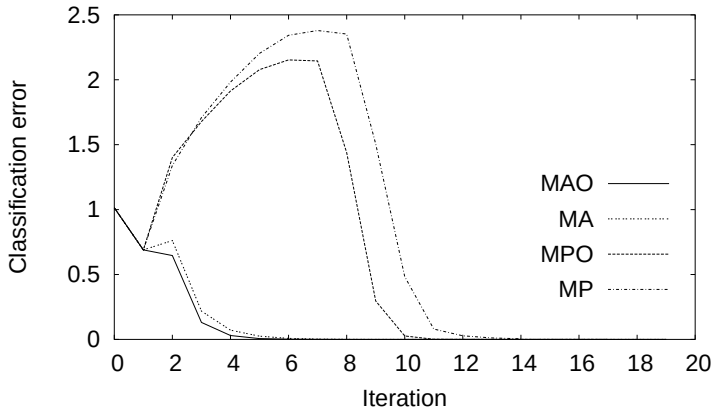

Figure 5: Evolution of the classification error obtained with the different selection strategies on the GT72 database.

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 "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).

The evolution of the mean precision during successive RF iterations (rounds) on the GT72 and GT100 databases are presented in Fig. 3 and 4. The "mean precision" value shown is obtained as the mean over all the images in the database of the precision measure described above. Clearly, the reduction of the redundancy between the images selected for labeling improves the results, both for MAO 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.

In some cases (GT9 and part of the classes in GT30) the MP and MPO criteria perform slightly better, during the first iterations, than the MA and MAO criteria. This behavior is typical for "trivial" classes, i.e. for which simple queries by example easily find all or most of the "relevant" images.

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 textual annotations 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. 5 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. As seen in Fig.

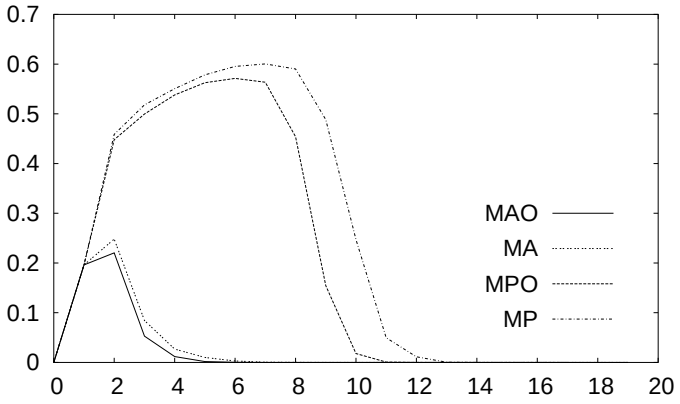

Figure 6: Evolution of the rate of "false positives" obtained with the different selection strategies on the GT72 database.

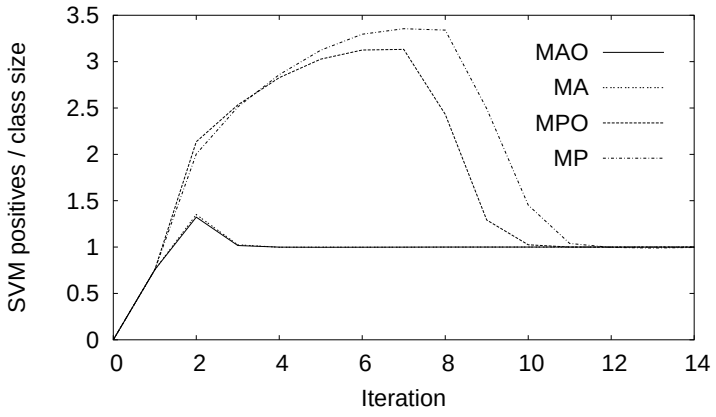

Figure 7: Evolution of the ratio between the images considered by the SVM as "relevant" and the number of images in a class for the different selection strategies on the GT72 database.

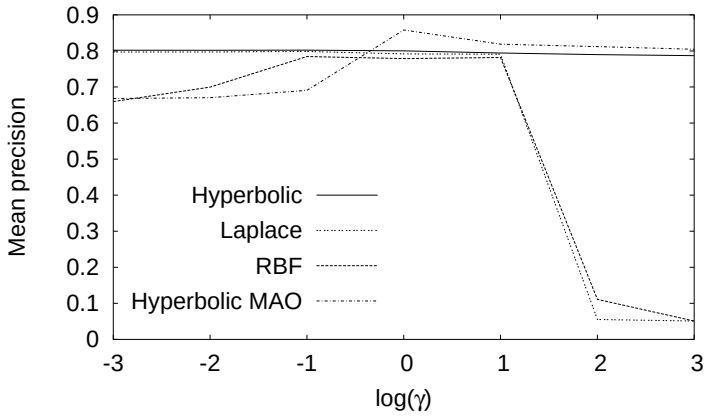

Figure 8: Sensitivity of the SVM to a scale parameter for the different kernels on the GT72 database.

6, a similar evolution is found for the rate of false positives, of primary importance for the extension of textual annotations.

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, for the different selection criteria on the GT72 database. The results obtained on the GT72 database are shown in Fig. 7. The convergence is significantly faster for the MAO and MA criteria, because MP and MPO do not focus on the frontier.

4 Invariance to scale and the choice of the kernel

During the study of several groundtruth databases we found that the size of the various classes often covers an important range of different scales (1 to 7 for the GT72 database, 1 to 5 for the GT100 database). We expect yet more significant changes in scale to occur from one database to another, from one user-defined class to another within a large database or between parts of the frontier of some classes. A too strong sensitivity of the learner to the scale of the data could then seriously 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 very classical kernel is highly sensitive to the scale parameter γ (the inverse of the variance of the Gaussian).

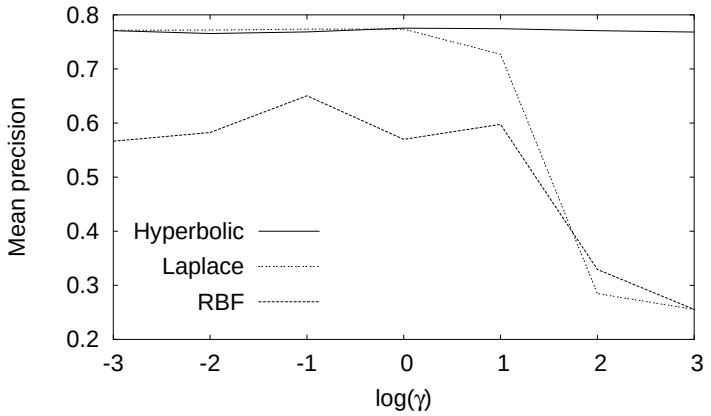

Figure 9: Sensitivity of the SVM to a scale parameter for the different kernels on the GT100 database.

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. 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). Note that 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. However, since the value of $K(x_i, x_j)$ decreases with an increase of the distance $d(x_i, x_j)$, our justification for the MAO criterion holds, as well as the justification of the MA criterion in [4]. Note that the use of a multiplicative parameter for the triangular kernel (e.g. $K(x_i, x_j) = -\gamma\|x_i - x_j\|$) has no effect on the SVM.

For all the kernels we used the L1 norm of the difference between descriptors because 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.

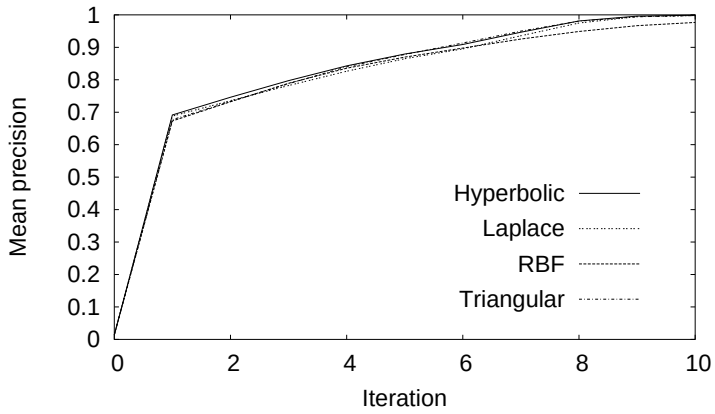

Figure 10: Comparison of the different kernels on the GT72 database, with the MP selection criterion.

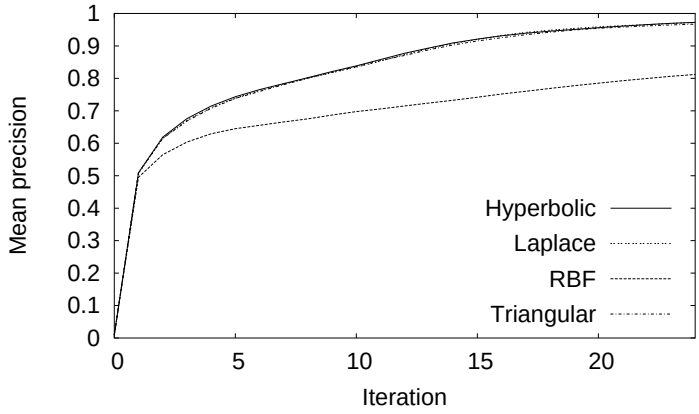

Figure 11: Comparison of the different kernels on the GT100 database, with the MP selection criterion.

The sensitivity of these kernels to the γ parameter with the MP strategy is shown in Fig. 8 for the GT72 database and in Fig. 9 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 strategy are shown in Fig. 10 and 11; for the Gaussian, Laplace and hyperbolic kernels, the scale parameters were set to their optimal values for the database.

From Fig. 8 and 9 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 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 easily 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. 8, this invariance is lost to some extent when the MAO selection criterion is employed.

For some groundtruth databases that do not show large differences in scale between classes, it may be possible to find optimal scale parameters for kernels such as Laplace or hyperbolic that will produce slightly better results than the triangular kernel. But in real applications, the scales of the user-defined classes cannot be known a priori and the scale parameter of a kernel cannot be adjusted online, so important variations can be expected for the performance of RF-based retrieval. The scale-invariance obtained by the use of the triangular kernel becomes then a highly desirable feature and makes this kernel a very good alternative.

5 Conclusion

In content-based image retrieval with relevance feedback, the strategy 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 strategy, based on a reduction of the redundancy between the images selected at every feedback round. By comparing this strategy to alternative strategies on several databases, we have shown that it performs better in ranking most of the "relevant" images before the others and also speeds up the convergence of the frontier around the class of interest. This last aspect is important when relevance feedback is used as a tool for the propagation of textual annotations.

By the study of several groundtruth databases we found significant changes in 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, the use of specific kernel functions such as the triangular kernel allows to obtain insensitivity to changes in scale while keeping performance at a good level.

References

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

[2] Nozha Boujemaa, Julien Fauqueur, Marin Ferecatu, François Fleuret, Valérie Gouet, Bertrand Le Saux, and Hichem 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] Klaus 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] Colin Campbell, Nello Cristianini, and Alexander 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] Edward Y. Chang, Beitao Li, Gang Wu, and Kingshy 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] Olivier 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] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[8] Ingemar J. Cox, Matthew L. Miller, Thomas P. Minka, Thomas Papathomas, and Peter 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] Ingemar J. Cox, Matthew L. Miller, Stephen M. Omohundro, and Peter 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, 1998.

[10] François Fleuret and Hichem 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] Theo Gevers and Arnold 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] Ralf Herbrich, Thore Graepel, and Colin Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, 2001.

[13] Feng Jing, Mingjing Li, Lei Zhang, Hong-Jiang Zhang, and Bo Zhang. Learning in region-based image retrieval. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2003.

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

[15] Bernhard Schölkopf and Alexander Smola. *Learning with Kernels*. MIT Press, 2002.

[16] Simon Tong and Edward 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] Simon Tong and Daphne 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] Vladimir Vapnik. *Estimation of dependencies based on empirical data*. Springer Verlag, 1982.

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

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

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399