

Relevance Feedback for Image Retrieval: a Short Survey *

Michel Crucianu, Marin Ferecatu, Nozha Boujemaa
INRIA Rocquencourt, B.P. 105
78153 Le Chesnay Cedex, France
{Michel.Crucianu, Marin.Ferecatu, Nozha.Boujemaa}@inria.fr

October 10, 2004

1 Introduction

The difficulty and 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 directly on their *visual content* (content-based image retrieval, CBIR) is of high interest today [16].

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. More 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 the descriptions that are meaningful for the users.

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 way of identifying what the user is looking for in the current retrieval session (the *target* of the user) is by including the user in the retrieval loop. For this, the session is divided

*in *State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction, Including Datamodels and Languages*, Report of the DELOS2 European Network of Excellence (6th Framework Programme)

into several consecutive *rounds*; at every round the user provides feedback regarding the retrieval results, e.g. by qualifying images returned as either “relevant” or “irrelevant” (*relevance feedback* or RF in the following); from this feedback, the engine learns the visual features of the images and returns improved results to the user. The RF mechanism implemented in a search engine should attempt to minimize the amount of interaction between the user and the engine required for reaching good results.

For other types of content, such as text or music, one can also find a gap between automatically extracted descriptions of content and meaningful retrieval criteria. In fact, RF was first introduced for the retrieval of text documents in [36] and [34]. The ease with which the relevance of an image can be evaluated and the persistent difficulty of dealing with the semantic gap in CBIR explains the rapid development of RF for image retrieval since the early work in [25], [30], [39], [35], [22] or [29].

We should note here that the semantic gap is not the unique explanation for the difficulties encountered in retrieval by content: the “numerical gap”, or the use of incomplete or confusing descriptions of the multimedia content, is an important complementary cause of problems in retrieval. To reduce this numerical gap, one should first attempt to find descriptors that are both rich and faithful.

In the following, we present the main issues related to relevance feedback for image retrieval and we review recent developments in this domain. We then mention a few promising research directions for the near future.

2 Objectives and formulation of the problem

Knowledge of the objectives of retrieval and of the characteristics of the data is important both for defining adequate RF mechanisms and for choosing appropriate evaluation methods for these mechanisms.

The first and most frequent objective consists in finding images that share some specific characteristic the user has in mind. The case of *target search* studied in [11] and [10], where the user is looking for that particular single image she has in mind, was further distinguished from the more common *category search*.

A complementary but less frequent use of RF was introduced in [27], [44] and consists in defining a class of images and extending textual annotations of some images in the class to the others.

2.1 Search for images sharing some characteristics

It is worth noting that this objective can be encountered in various scenarios, corresponding to different expectations of the user, such as:

- *Explore* and search for *some* “relevant” items. In this case, the user has a rather vague prior notion of relevance and relies on the exploration of the image base to clarify it. The search engine must allow for a rather unfocused exploration, but it doesn’t need to find all the “relevant” images, nor to exclude all the “irrelevant” images, since the user will just pick some “relevant” ones in the end.
- Retrieve *most* items in a set of “relevant” ones. In this case, the user would like to find all or most of the images that share some specific characteristic she has in mind. The exploratory behaviour is more focused here, but again the search engine doesn’t need to exclude all the “irrelevant” images, since the user will ignore the few that may remain in the end.

Relevance is usually defined by a characteristic that is shared by some images. It can be a perceptual characteristic or a more semantic one, and it may concern entire images or parts of images.

This objective is usually considered to correspond to a ranking problem, where the images must be ordered and returned to the user by decreasing relevance. However, it is generally accepted that a precise ranking of the “relevant” or of the “irrelevant” images is not required, partly because the user may be unable to choose between two alternative rankings. The search engine should simply rank most of the “relevant” images before the “irrelevant” ones.

Setting up a frontier between “relevant” and “irrelevant” images is not important here, so this is not a classification problem. Estimating a density function for the “relevant” images is not necessary either, since we are only interested in the high density regions of the image description space (an estimation of the density would also be unreliable with the few examples that we can expect). We can also consider that the search for images sharing some characteristics is a problem of *identifying the modes* of a distribution.

To evaluate the performance of an RF mechanism attempting to solve such a retrieval problem, it is necessary to measure the quality of the ranking of “relevant” images before the “irrelevant” ones and the speed of improvement of this ranking during successive feedback rounds.

2.2 Discrimination of a class of images

The extension of textual annotations of some images to the others in a same class should help reducing the number of images that have to be manually annotated. For this procedure to be efficient, the effort required for defining the class corresponding to the annotation should be much lower than the effort for a manual annotation. This is why it was suggested in [27], [44] to use RF to interactively define the class of images corresponding to an annotation. This is clearly a classification problem: the frontier between the images belonging to the target class (the “relevant” images in this case) and the others (the “irrelevant” images) must be reliably identified. A ranking of the “relevant” and “irrelevant” images is not sufficient for obtaining such a classification.

The rate of false positives (images that do not belong to the class but are assigned to it and, consequently, receive a wrong annotation) is considered to be more important than the rate of false negatives (images that belong to the class but are not assigned to it and do not receive the annotation). However, the frontier need not always be crisp and degrees of confidence can be associated to the resulting annotation.

For this objective, it is the classification performance of RF that must be evaluated: one should measure the classification error (or only the rate of false positives) and the speed of reduction of this error during successive feedback rounds.

2.3 Image representation

The representation of individual images also has an impact on the RF mechanism employed. In existing work on the CBIR with RF, two different representation schemes were used for the images:

- Most of the time, the global visual appearance of the images is described using a combination of global signatures including colour, texture and shape information. Images are then represented by fixed-length vectors in a *description space*.
- In some publications, such as [33], [20], [23] or [24], an image is considered to be a set (or a “bag”) of regions obtained by an automatic segmentation. Every region can be described by colour, texture and shape. Additionally, some information regarding the configuration of the regions can be available. An image is then represented as variable-length collection of region signatures, possibly including configuration information. From user feedback concerning entire images, the search

engine is also expected to learn (in a variant of *multiple instance learning*) which regions are important for the current search session and which regions to ignore. However, the difficulty of this problem would be significantly reduced if the user could select and provide feedback for individual regions.

2.4 General assumptions

One can customize an RF mechanism if one knows the characteristics of the scenario, of the target application and of its users. It is nevertheless important to remind some general assumptions that are usually made when developing RF mechanisms for CBIR:

1. The discrimination between “relevant” and “irrelevant” images must be *possible* with the available image descriptors.
2. There is some *relatively simple* relation between the topology of the description space and the characteristic shared by the images the user is searching for.
3. “Relevant” images are a *small part* of the entire image database.
4. While part of the early work on RF assumed that the user could (and would be willing to) provide a rather rich feedback, including “relevance notes” for many images, the current assumption is that this feedback information is scarce: the user will only mark a few “relevant” images as positive and some very different images as negative.

2.5 Sources and nature of the available information

Given the rather small amount of interaction with the user during an RF session, it is important to use all the available information to improve the retrieval results. In Tab. 1 we show what sources of information one can try to exploit. Few existing publications address this important issue (see e.g. [3] and [15]).

3 Relevance feedback mechanisms

To preserve interactivity, the RF mechanism implemented in a search engine must operate in real time. It is expected to maximize the ratio between the quality of the retrieval results and the amount of interaction between the user and the system.

Table 1: Sources and nature of the information that can be exploited in relevance feedback

Source	Time period		
	other sessions	current session	current round
prior	domain-specific similarity, prior clustering	nature and context of session	-
other users	retrieval-based correlations	-	-
User	model of subjective perceived similarity	answers in previous rounds	answer in current round

An RF mechanism has two components: a learner and a selector. At every feedback round, the user marks (part of) the images returned by the search engine as “relevant” or “irrelevant”. The learner exploits this information to re-estimate the target of the user. With the current estimation of the target, the selector chooses other images that are displayed by the interface of the search engine; the user is asked to provide feedback on these images during the next round. In the following, we briefly present the evolution of the learners and of the selection criteria in the recent years.

3.1 Learners

In RF, the learner must use the training data, i.e. the images marked by the user during subsequent feedback rounds, and sometimes prior knowledge (see Tab. 1) in order to estimate the target of the user.

3.1.1 Difficulty of learning in relevance feedback

The task of the learner is particularly difficult in the context of RF for several reasons (see also [6], [47]):

- The amount of training data is very low, usually much lower than the number of dimensions of the description space.
- There are usually much fewer positive examples (images marked by the user as “relevant”) than negative examples (images marked as “relevant”). The learner must have a low sensitivity to this imbalance in the training set or some corrective must be found.

- The target class may have a rather complex shape or even several, rather disconnected modes. Together with the fact that training data is scarce, this can severely limit the generalization we can expect.
- To preserve interactivity, both learning from the training examples and the evaluation of the remaining images according to the selection criterion must be very fast. The computation cost can then be a very important criterion in the choice of a learning method.

3.1.2 Evolution of learners for relevance feedback

Early approaches to RF for CBIR are still close to the classical query by visual example (QBVE) framework: they assume the existence of an ideal query point that, if found, would provide the appropriate answer to the user when used for QBVE. These approaches belong to the family of “query point movement” (QPM) methods, for which the task of the learner consists in finding, at every round, a better query point together with a re-weighting of the individual dimensions of the description space. Learning can rely on the positive examples alone, such as for the re-weighting scheme in [39] or the Mahalanobis distance-based proposal in [22], or on both positive and negative examples, such as the methods put forward in [35], [29] or [32].

All these methods make the strong assumption that the target class has an elliptical shape, but some go even further by considering that the axes of the ellipsoid are the original axes of the description space (i.e. that the covariance matrix of the target class is diagonal). Learning corresponds here to the estimation of the parameters of a Gaussian distribution.

To our knowledge, these strong assumptions regarding the shape of the target class were first removed in [28] and [31]. The *query expansion* scheme put forward in [31] consists in performing an online clustering of the examples and evaluating all the other images using a nearest-neighbour decision with respect to these clusters. In [28] the density of the positive examples and the density of the negative examples are estimated with a Parzen window method; the decision function used for ranking the images returned to the user is the difference between the two densities (the first minus the second). The additiveness of this density estimation method makes it incremental, i.e. at every feedback round a fixed number of terms is added to the decision function; this is very important for the cost-effectiveness of the algorithm.

Recent work on RF often relies on support vector machines (SVM, [42], [38]). With SVMs, the data is usually first mapped to a higher-dimensional feature space using a non-linear transform associated to a reproducing kernel; linear discrimination between classes is then performed in this feature space;

the frontier only takes into account the support vectors, which are (loosely speaking) those examples that are closest to the frontier. Learning is based on constrained quadratic optimisation. The reader should refer to e.g. [38] for a detailed presentation of SVMs and other kernel methods.

In RF, SVMs appear to be the learners of choice for several reasons:

- The decision function of an SVM allows both the definition of a frontier and the ranking of images.
- For most choices of the kernel, SVMs avoid too restrictive assumptions regarding the data (e.g. the shape of the target class).
- SVMs are very flexible. For example, prior knowledge regarding the problem can be used to tune the kernel.
- SVMs allow fast learning (with the rather limited number of examples provided by feedback) and relatively fast evaluation for medium-sized databases.
- By relying only on support vectors, SVMs are usually less sensitive than density-based learners to the imbalance between positive and negative examples in the training data.

It should be noted, however, that SVMs lack (in their original formulation) the incremental character and the corresponding cost-effectiveness of Parzen density estimation.

While most of the existing work using SVMs for RF concentrates on 2-class SVMs (see [21], [46], [40], [23] or [24] to mention only a few) that must learn to discriminate positive and negative examples, 1-class SVMs were also put forward in [8] in order to learn from positive examples only. 1-class SVMs are able to estimate the support of the distribution of positive examples (images marked by the user as “relevant”). Beside this ability to use only positive examples, another (rather implicit) argument in favour of 1-class SVMs (see [8]) is that 2-class SVMs tend to overestimate the target class. However, by completely ignoring the negative examples when they exist, in many cases 1-class SVMs cannot avoid including in the images returned to the user many “irrelevant” images.

Another kernel method was suggested in [45] and [46] as a learner for RF: kernel biased discriminant analysis (kernel BDA). Linear BDA identifies the dimensions of the description space according to their effectiveness in discriminating between positive and the other examples. Kernel BDA does the same, but in the feature space associated to the kernel rather than in the original description space. In [46] kernel BDA is found to produce better

results than both 2-class SVMs and kernel Fisher discriminant analysis (see [38]). However, since kernel BDA is a “mass”-based method (like standard discriminant analysis), we expect it to be sensitive to the imbalance between positive and negative examples, and to encounter difficulties when very few examples are available.

Various kernels were used for the SVM classifiers in RF. The linear kernel, $K(x_i, x_j) = x_i x_j$, can only be used when a linear discrimination between classes can be expected in the original description space, which is generally not the case for image signatures.

A well-known and often employed kernel is the Gaussian (or Radial Basis Function) kernel, $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$. However, this kernel is highly sensitive to the scale parameter γ (the inverse of the variance of the Gaussian).

The use of the Laplace (exponential) kernel, $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|)$, was advocated in [7] for histogram-based image descriptors. In [24], this kernel was also found to provide better results than the Gaussian kernel in CBIR with RF.

The hyperbolic kernel, $K(x_i, x_j) = 1/(\varepsilon + \gamma\|x_i - x_j\|)$, can be computed fast and was recently evaluated for RF with rather good results [13]. The scale parameter is again γ (ε translates into a multiplicative constant plus a change in γ and is only used to avoid numerical problems).

All the kernels we mentioned up to now are positive definite kernels. The triangular kernel, $K(x_i, x_j) = -\|x_i - x_j\|$, was introduced in [4] as a *conditionally* positive definite kernel, but the convergence of SVMs remains guaranteed with this kernel [37]. In [14] 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 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.

By the study of several groundtruth databases, it was found in [13] that the size of the various classes often covers an important range of different scales; yet more significant changes in scale are expected to occur from one user-defined class to another within a large database in real-life RF applications. 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 comparison performed in [13] shows that the triangular kernel outperforms the other kernels we mentioned, with the Gaussian kernel producing the lowest performance. Indeed, 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 and hyperbolic kernels, an increase of γ beyond 1 has some impact on the results (this impact is stronger for the Laplace kernel), while a reduction of γ does not have significant consequences. This is easily explained by the fact that for small γ the Laplace and hyperbolic kernel become similar to the triangular kernel. Since 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, the scale-invariance obtained by the use of the triangular kernel is a highly desirable feature

We must note that the norm $\|x_i - x_j\|$ used when computing the kernel has an impact on the results: the L_1 norm outperforms the L_2 norm. It was found in [6] that the use of the perceptual dissimilarity function defined in [26] instead of $\|x_i - x_j\|$ improves retrieval results. This perceptual dissimilarity, which is not a metric, can be obtained in the following way: to find the dissimilarity between two vectors, only the n dimensions corresponding to the smallest absolute differences are retained (n is fixed for a database and identified by trial and error). This perceptual “distance” was justified using psychological considerations and was experimentally found in [26] to work better than L_1 , L_2 and L_p with $0 < p < 1$ (this fractional measure, put forward in [1], is not a metric either).

When images are represented as variable-length collections of region signatures, [24] suggests the use of the Earth Mover’s Distance (EMD) instead of $\|x_i - x_j\|$ when computing the Gaussian kernel. Unfortunately, the computation cost is high for the EMD and the resulting kernel does not necessarily satisfy the conditions that guarantee convergence of the SVM. While this new kernel is found to work better than the standard Gaussian kernel and the Laplace kernel, the comparison is biased by the fact that the image representation is finer when the variable-length representation (only exploitable with the new kernel) is employed.

3.2 Selection criteria

In much of the work on RF, the images for which the user is asked to provide feedback at the next round are simply those that are currently considered by the learner as potentially the most “relevant”; also, in a few cases these

images were randomly selected. It is important to understand the goals of the selection criterion and how it can be improved.

3.2.1 Goals of the selector

An RF session is open-ended, since the user is supposed to be able to provide feedback any time during the session. It is then difficult to define two distinct stages in the session: the identification of the target class, followed by the presentation of its images to the user. As a consequence, the selection strategy has two different and potentially conflicting goals during each feedback round:

1. Given the current state of knowledge of the learner, provide the user with as many “relevant” images as possible.
2. Elicit from the user as much information as possible regarding the distinction between “relevant” and “irrelevant” (maximize the transfer of information from the user to the system).

3.2.2 Select the “most positive” images

The “return the most positive images” criterion (MP in the following) selects those images that are currently considered by the learner as the most “relevant”. This strategy is the most frequently found in the literature on RF and focuses on the first goal just mentioned. It has the advantage that the user gets quite early many items from (or close to) the target class, and consequently some satisfaction, but the disadvantage that a more or less complete identification of the target class may take longer.

3.2.3 Select the “most informative” images

A “return the most informative images” criterion (MI in the following) focuses on the second goal mentioned above. Valuable ideas were introduced in [11] and [10], where the problem under focus is *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 single 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 remove a maximal amount of uncertainty regarding the target. 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. The entropic criterion employed

in [11], [10] does not scale well to *category* search or to the selection of more than 2 images. Computational optimisations must be found, relying on the use of specific learners and, possibly, specific search contexts.

Based on the definition of *active learning* (see [2], [9]), the selection of examples for training SVMs to perform general classification tasks is studied in [5]. 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 sub-optimal, so the fastest reduction in error can be achieved by selecting the example that is closest to the current estimation of the frontier. Following to the classical formulation of active learning, the authors only consider the selection of single examples for labelling (for addition to the training set) at every round.

For SVM learners too, several selection criteria are presented in [41] and applied to the classification of texts. 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 [41] 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 labelled 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. In order to minimize the number of learning rounds, the user is asked to label several examples at every round and these examples are all selected according to the MA criterion. In [40] 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 labelling.

Note that the MA criterion in [41], [40] is the same as the one put forward in [5] 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 [19] instead of SVMs or the more sophisticated criteria put forward in [41], 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 may sometimes produce a faster convergence of part of the frontier during the first few rounds of RF.

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).

It was recently suggested in [13] 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 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 $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$). This selection criterion was called “most ambiguous and orthogonal” (MAO) and is considered to be a better approximation to MI.

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.

Since the triangular kernel is not positive definite but only conditionally positive definite, the account provided in [41] 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 $d(x_i, x_j)$, the justification for the MAO criterion in [13] does hold, as well as the justification of the MA criterion in [5].

3.2.4 Hybrid selection strategies

Since MP and MI focus each on a single goal, a hybrid selection strategy may be able to find a good compromise. The hybrid strategy suggested

in [43] for text retrieval consists in using the MP criterion for k texts and the MA criterion for the remaining $n - k$ texts that are selected at every feedback round. Also, k increases over time during the interactive retrieval session: more ambiguous examples are presented in the beginning, to identify the target class faster, while later the ratio of (potentially) “relevant” texts increases. For the databases considered in [43], this hybrid strategy compares well to the use of the MP or MA selection criteria. Further comparisons should be performed to compare the superiority of this strategy and to find how to choose the parameters of the transition from more MA to more MP.

3.3 Structure of the session

The initialisation of search with RF is a very important issue. A good starting point can be provided by the user (example image), found with the help of visual summaries, or identified by the logical composition of categories from a visual thesaurus [12]. If such a good starting point is not available, then feedback must be used both for finding some “relevant” image starting from a random initial sample (exploratory stage) and then for retrieving further “relevant” images (“exploitation” stage). An exploratory behaviour must also be supported because some target classes have several distinct modes. Note that in such cases the significance of feedback changes during the session: in exploratory stages the user will mark as positive images that have some similarity with target images but cannot be considered as “relevant”, while in exploitation stages she is expected to mark as positive only “relevant” images. While the exploratory behaviour was addressed in early work on QPM methods, most recent work focuses on the exploitation stage.

4 Evaluation of search with relevance feedback

4.1 How to evaluate

Evaluating relatively general improvements to RF mechanisms 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 a user would 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.

The policy of the user when providing feedback can have a strong impact on the evaluation of RF. In most cases, the user is expected to mark as either positive or negative all the images selected by the RF mechanism as candidates for feedback. This assumes a “stoic” user, which is not very realistic. The behaviour of RF mechanisms in the presence of more evolved user policies or of more typical policies (e.g. mark only a few of the images returned or make some mistakes in marking) should be further studied.

4.2 What to measure

Since user satisfaction is very subjective and experimenting with users is difficult, other performance measures were defined, relying on the use of groundtruth databases for the evaluation of retrieval.

The performance measures usually employed consist in the evaluation of the proportion of “relevant” images in the top N returned by the search engine (N being the number of images in the target class of the groundtruth database). The evolution of this measure during successive feedback rounds is an indication of the speed of convergence to the target ranking. We can also mention here the use of precision vs. recall after some fixed number of feedback rounds. Note that these performance measures are appropriate for the first objective of search with RF, described in Sect. 2.1, and not for the second, described in Sect. 2.2. To evaluate performance with respect to this second objective, one should rather measure the evolution of the classification error or of the rate of false positives during successive feedback rounds (see [13]).

5 Directions in retrieval with relevance feedback

Among the important issues that deserve more effort in the near future we can mention:

- The prior information available (see Tab. 1) must be better exploited by the RF mechanisms. This should produce a general improvement of the retrieval performance of RF.
- The impact of the characteristics of the data and of the policy of the user on both the learner and the selector must be addressed. This should allow an improvement of RF mechanisms used in specific but maybe frequent settings.
- The scaling of RF to very large image databases is an important issue that was not extensively studied. Recent approaches, such as [17] and [18], are promising.

References

- [1] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Proceedings of the 8th International Conference on Database Theory*, pages 420–434, 2001.
- [2] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.
- [3] Ilaria Bartolini, Paolo Ciaccia, and Florian Waas. Feedbackbypass: A new approach to interactive similarity query processing. In *Proceedings of the 27th International Conference on Very Large Data Bases*, pages 201–210, September 2001.
- [4] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, 1984.
- [5] 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.

- [6] Edward Y. Chang, Beita 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.
- [7] Olivier Chapelle, P. Haffner, and Vladimir N. Vapnik. Support-vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.
- [8] Yunqiang Chen, Xiang Sean Zhou, and Thomas S. Huang. One-class SVM for learning in image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'01)*, 2001.
- [9] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- [10] Ingemar J. Cox, Matthew L. Miller, Thomas P. Minka, Thomas Pappathomas, 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.
- [11] 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. IEEE Computer Society, 1998.
- [12] Julien Fauqueur and Nozha Boujemaa. New image retrieval paradigm: Logical composition of region categories. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'2003)*, September 2003.
- [13] Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa. Reducing the redundancy in the selection of samples for SVM-based relevance feedback, May 2004.
- [14] 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.
- [15] Jérôme Fournier and Matthieu Cord. Long-term similarity learning in content-based image retrieval. In *Proceedings of the International Conference on Image Processing*, pages 441–444, 2002.

- [16] 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.
- [17] Douglas R. Heisterkamp and Jing Peng. Kernel indexing for relevance feedback image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'03)*, 2003.
- [18] Douglas R. Heisterkamp and Jing Peng. Kernel VA-files for relevance feedback retrieval. In *Proceedings of the first ACM international workshop on Multimedia databases*, pages 48–54. ACM Press, 2003.
- [19] Ralf Herbrich, Thore Graepel, and Colin Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, 2001.
- [20] Pengyu Hong and Thomas S. Huang. Spatial pattern discovering by learning the isomorphic subgraph from multiple attributed relation graphs. In Sébastien Fourey, Gabor T. Herman, and T. Yung Kong, editors, *Electronic Notes in Theoretical Computer Science*, volume 46. Elsevier, 2001.
- [21] Pengyu Hong, Qi Tian, and Thomas S. Huang. Incorporate support vector machines to content-based image retrieval with relevant feedback. In *Proceedings of the 7th IEEE International Conference on Image Processing*, September 2000.
- [22] Yoshiharu Ishikawa, Ravishankar Subramanya, and Christos Faloutsos. Mindreader: Querying databases through multiple examples. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, pages 218–227. Morgan Kaufmann Publishers Inc., 1998.
- [23] Feng Jing, Mingjing Li, Hong-Jiang Zhang, and Bo Zhang. Learning region weighting from relevance feedback in image retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002.
- [24] 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.
- [25] T. Kurita and T. Kato. Learning of personal visual impression for image database systems. In *Second Intl. Conf. on Document Analysis and Recognition*, pages 547–552, 1993.

- [26] Beita Li, Edward Chang, and Yi Wu. Discovery of a perceptual distance function for measuring image similarity. *MultiMedia Systems*, 8:512–522, 2003.
- [27] Ye Lu, Chunhui Hu, Xingquan Zhu, Hong-Jiang Zhang, and Qiang Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proceedings of the 8th ACM International Conference on Multimedia*, pages 31–37. ACM Press, 2000.
- [28] Christophe Meilhac and Chahab Nastar. Relevance feedback and category search in image databases. In *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, pages 512–517. IEEE Computer Society, June 1999.
- [29] Chahab Nastar, Matthias Mitschke, and Christophe Meilhac. Efficient query refinement for image retrieval. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 547–552. IEEE Computer Society, 1998.
- [30] Rosalind W. Picard, Thomas P. Minka, and Martin Szummer. Modeling user subjectivity in image libraries. In *IEEE Int. Conf. On Image Processing*, pages 777–780, 1996.
- [31] Kriengkrai Porkaew and Kaushik Chakrabarti. Query refinement for multimedia similarity retrieval in mars. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 235–238. ACM Press, 1999.
- [32] Kriengkrai Porkaew, Michael Ortega, and Sharad Mehrotra. Query reformulation for content based multimedia retrieval in MARS. In *ICMCS, Vol. 2*, pages 747–751, 1999.
- [33] Aparna Lakshmi Ratan, Oded Maro, W. Eric, L. Grimson, and Tomás Lozano-Pérez. A framework for learning query concepts in image classification. In *Proceedings of CVPR'99*, pages 1423–1429, 1999.
- [34] C. J. van Rijsbergen. *Information retrieval*. Butterworths, London, 2 edition, 1979.
- [35] Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool in interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

- [36] Gerard Salton. *Automatic Information Organization and Retrieval*. McGraw-Hill, 1968.
- [37] Bernhard Schölkopf. The kernel trick for distances. In *Advances in Neural Information Processing Systems*, volume 12, pages 301–307. MIT Press, 2000.
- [38] Bernhard Schölkopf and Alexander Smola. *Learning with Kernels*. MIT Press, 2002.
- [39] Stan Sclaroff, Leonid Taycher, and Marco La Cascia. Imagerover: A content-based image browser for the world wide web. In *Proceedings of the 1997 Workshop on Content-Based Access of Image and Video Libraries (CBAIVL'97)*, page 2. IEEE Computer Society, 1997.
- [40] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *Proceedings of the 9th ACM international conference on Multimedia*, pages 107–118. ACM Press, 2001.
- [41] 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.
- [42] Vladimir Vapnik. *Estimation of dependencies based on empirical data*. Springer Verlag, 1982.
- [43] Zhao Xu, Xiaowei Xu, Kai Yu, and Volker Tresp. A hybrid relevance-feedback approach to text retrieval. In *Proceedings of the 25th European Conference on Information Retrieval Research, Lecture Notes in Computer Science*, volume 2633. Springer-Verlag, April 2003.
- [44] Hong-Jiang Zhang and Z. Su. Improving CBIR by semantic propagation and cross-mode query expansion. In *Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MM-CBIR'01)*, pages 83–86, September 2001.
- [45] Xiang Sean Zhou and Thomas S. Huang. Comparing discriminating transformations and SVM for learning during multimedia retrieval. In *Proceedings of the 9th ACM international conference on Multimedia*, pages 137–146. ACM Press, 2001.
- [46] Xiang Sean Zhou and Thomas S. Huang. Small sample learning during multimedia retrieval using BiasMap. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.

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