An exploration of diversified user strategies for image retrieval with relevance feedback

Michel Crucianu, Jean Philippe Tarel, Marin Ferecatu

An Exploration of Diversified User Strategies for Image Retrieval with Relevance Feedback

Michel Crucianu [a,b] Jean-Philippe Tarel [a,c] Marin Ferecatu [a]

[a] *INRIA Rocquencourt, 78153 Le Chesnay Cedex, France*

[b] *Vertigo-CEDRIC, CNAM, 292 rue St. Martin, 75141 Paris Cedex 03, France*

[c] *Laboratoire Central des Ponts et Chaussées, 58 Bd. Lefebvre, 75015 Paris, France*

Abstract

Given the difficulty of setting up large-scale experiments with real users, the comparison of content-based image retrieval methods using relevance feedback usually relies on the emulation of the user, following a single, well-prescribed strategy. Since the behavior of real users cannot be expected to comply to strict specifications, it is very important to evaluate the sensitiveness of the retrieval results to likely variations of users' behavior. It is also important to find out whether some strategies help the system to perform consistently better, so as to promote their use. Two selection algorithms for relevance feedback based on support vector machines are compared here. In these experiments, the user is emulated according to eight significantly different strategies on four ground truth databases of different complexity. It is first found that the ranking of the two algorithms does not depend much on the selected strategy. Also, the ranking of the strategies appears to be relatively independent of the complexity of the ground truth databases, which allows to identify desirable characteristics in the behavior of the user.

Key words: content-based retrieval, user strategy, user emulation, evaluation

1 Introduction

The scarcity and inherent incompleteness of the textual annotations of multimedia content promote the use of search by content in multimedia databases [1], in spite of limitations due to the "semantic gap". To go beyond simple query by example in order to identify more precisely what a user is actually looking for, search engines must include the user in the retrieval loop.

In search with *relevance feedback*, a session is divided into several consecutive rounds during which the user provides feedback regarding the retrieved

results, usually by qualifying content items returned as either "relevant" or "irrelevant". From this feedback, the engine *learns* the features associated with the desired content and proposes to the user the newly retrieved results. The many relevance feedback methods developed, mostly in the content-based image retrieval community, endeavor to minimize the amount of interaction required for ranking most of the "relevant" images before "irrelevant" ones.

Since large-scale experiments with real users are costly and difficult to set up, evaluations and comparisons of relevance feedback algorithms usually rely on the use of ground truth databases and on an emulation of the user. Such a database is partitioned into well-defined classes of images and the emulated user follows a single, well-prescribed strategy in qualifying returned images as "relevant" or "irrelevant".

But the behavior of *real* users in qualifying the returned images cannot be expected to comply to strict specifications. Moreover, it is reasonable to imagine that the choice of a strategy has an impact on the quality of the results. How general and meaningful are then the conclusions drawn from comparisons performed with ground truth databases? This is the main issue addressed in the following. For similar cost reasons, we also emulate the user behavior and rely on ground truth databases. However, in the evaluation presented below, *multiple* user strategies are employed, as well as *several* image databases of *different* complexity.

While such an evaluation cannot replace large-scale experiments with real users, it allows to explore the impact of various user strategies, at a lower cost and in a controlled way. This study should bring more confidence in the comparisons between relevance feedback algorithms and to provide some insight into possible relations between the algorithm, the database and the user strategy.

It is also important to find out whether some user strategies help the system perform consistently better, or provide more robustness to changes in the complexity of the database. Such a strategy can then be recommended to the users, even if they would not follow it strictly.

The next section provides details about the relevance feedback algorithms compared here. The eight user strategies that are investigated are introduced in Section 3. Section 4 presents the setting of the study and includes a description of the four ground truth databases employed. The results of all the comparative evaluations are presented and discussed in Section 5.

2 Relevance Feedback with Support Vector Machines

It is assumed that every image is represented by a signature describing its visual content. The image signatures employed here are presented in Section 4.1. A relevance feedback method is defined by two components: a *learner* and a *selector*. At every feedback round, the learner uses the signatures of the images labeled as "relevant" or "irrelevant" by the user to re-estimate a split of the signature space in "relevant" and "irrelevant" regions. Given the current estimation of this split, the selector chooses according to its selection criterion the images for which the user is asked to provide feedback at the next round.

Much work on relevance feedback relies on the use of support vector machines [2] to discriminate between "relevant" and "irrelevant" images (e.g. [3–5]). A support vector machine maps the data (image signatures in the case under study) to a higher-dimensional feature space using a non-linear transformation associated to a reproducing kernel; then, it performs linear discrimination between "relevant" and "irrelevant" items in this feature space. The discriminating hyperplane is only defined by the *support vectors* and learning is based on quadratic optimization under linear constraints. Learning leads to a decision function over the space of signatures. For every signature, the value of this function is the signed distance between the hyperplane and the mapping of the signature in the higher-dimensional feature space.

The decision function can be used for ranking all the images in the database: the most "relevant" images are those for which the decision function takes the highest positive values. Among the advantages of support vector machines over other learners in the relevance feedback context, can be mentioned the absence of too restrictive assumptions regarding the data, the flexibility (can be tuned by kernel engineering) and the fast learning and evaluation for medium-sized databases.

Most studies consider the Gaussian kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$, with a fixed value for the scale parameter γ. The high sensitivity of the Gaussian kernel to the scale parameter is an important drawback for relevance feedback [6]. Indeed, since significant variations in spatial scale from one class to another can be found for classes in ground truth databases (as for user-defined classes in real-world applications), *any* fixed value for the scale parameter will be inadequate for many of these classes. Following [7,6], the *conditionally* positive definite "angular" kernel $K(\mathbf{x}_i, \mathbf{x}_j) = -\|\mathbf{x}_i - \mathbf{x}_j\|$ is employed here. It was shown in [8] that the convergence of support vector machines using this kernel is nevertheless guaranteed. In [7], the angular kernel was found to have the interesting property of making the support vector machine insensitive to the scale of the data (within the limits set by the regularization bound C).

In most systems, the selection consists in choosing the images currently considered by the learner to be the most relevant. This criterion will be called the selection of the "Most Positive" (MP) candidates. An *active learning* framework for relevance feedback using support vector machines was introduced in [9,10]. The associated selection criterion consists in choosing the images whose signatures are the closest to the current frontier between "relevant" and "irrelevant" images. This will be called the selection of the "Most Ambiguous" (MA) candidates. A drawback of the MA criterion is that very similar images may be selected together. An additional condition of low redundancy was put forward in [6] and requires the selection of candidates that are far apart, in order to better explore the current frontier between "relevant" and "irrelevant". More specifically, consider that $\mathbf{x}_i$ and $\mathbf{x}_j$ are the signatures of two candidate images. To have $\mathbf{x}_i$ and $\mathbf{x}_j$ far apart, a low value for $K(\mathbf{x}_i, \mathbf{x}_j)$ is required, since the value of the angular kernel decreases with an increase of the distance $d(\mathbf{x}_i, \mathbf{x}_j)$. The inclusion of this low redundancy condition in the MA criterion will be denoted by MAO (Most Ambiguous and Orthogonal).

To implement the MAO criterion, a larger set of unlabeled images is first selected using MA. Then, the MAO selection is obtained by iteratively choosing as a new candidate the vector $\mathbf{x}_j$ that minimizes the highest value of $K(\mathbf{x}_i, \mathbf{x}_j)$ for all $\mathbf{x}_i$ already included in the current MAO selection: $\mathbf{x}_j = \text{argmin}_{\mathbf{x} \in S} \max_i K(\mathbf{x}, \mathbf{x}_i)$. S is the set of images selected by MA and not yet included in the MAO selection, while $\mathbf{x}_i$ are the images already in the MAO selection. The number of unlabeled images preselected with MA is a multiple of the number of images for which the user is asked to provide feedback at the next round ("window size", ws below). The size of the preselection is $2 \times ws$ because in previous experiments [6] it was found to be a good compromise between ambiguousness and low redundancy. The MP and MAO selection criteria are compared in the following, using support vector machines with the angular kernel.

3 User Strategies

The evaluation and the comparison of relevance feedback algorithms usually rely on an emulation of the user according to the following strategy: given a target class from a ground truth database, the user qualifies at every round *all* the images returned by the selector as either "relevant" (belong to the target class) or "irrelevant" (don't belong to the target class) and makes no mistakes; this will be called "stoic" user behavior (STO below).

Given the cost and difficulty of setting up large-scale experiments with real users, ground truth databases and user emulation are also employed, but variations of the behavior of the users are investigated (in a controlled way) by

defining the following seven new strategies:

(1) An "annoyed" user (ANN) labels only a fixed ratio (50% in the experiments reported here) of the images returned by the selector; the user randomly chooses the images to label, but makes no mistake when labeling them.
(2) A "greedy" user (GRE) correctly labels all the "relevant" images (images belonging to the target class), if present, together with the most "irrelevant" image (not belonging to the target class), if any is present.
(3) A "cooperative" user (COO) correctly labels the most "relevant" image if at least one is present and the most "irrelevant" image if not.
(4) A "minimalist" user (MIN) correctly labels one randomly chosen "relevant" image, if present, and one randomly chosen "irrelevant" image, if present.
(5) An "optimistic" user (OPT) correctly labels the most ambiguous among the "relevant" images and the most "irrelevant" image.
(6) A "restrictive" user (RST) correctly labels the most "relevant" image if at least one is present and all the "irrelevant" images.
(7) A "tired" user (TIR) labels all the images returned by the selector, but makes mistakes (i.e. labels as "irrelevant" a "relevant" image, or as "relevant" an "irrelevant" image) with a given probability (of 0.1 in the following).

While other strategies can be defined, we consider that those listed above cover the main variations expected for the behavior of the users: reliable vs. error-prone, diligent vs. uninvolved, and different ways of collaborating with the system. Most of these strategies were identified by studying the retrieval sessions of a few knowledgeable but non professional users. Two strategies, GRE and RST, were specifically included to study the benefits of labeling many more "relevant" or many more "irrelevant" images. Real behaviors appear to follow a mixture of several strategies, where MIN or ANN steps are more frequent than STO steps.

Four of these strategies (GRE, COO, OPT and RST) rely on the evaluation of a "degree of relevance" by the user. Since the crisp classes of typical ground truth databases do not provide such information, an attempt was made to obtain it automatically. For every class of the ground truth, a support vector machine is trained using as positive examples all the images of the class and as negative examples the rest of the images; the decision function of this support vector machine is then employed for emulating the user during the retrieval sessions. Given a target class and a set of images shown to the emulated user, the more positive the value of this function is, the higher the "relevance", and the more negative its value is, the higher the "irrelevance"; the most ambiguous "relevant" image (for OPT) is the one for which the value of the function is positive but closest to 0. The evaluation of the degree of relevance

by real users was found to be relatively subjective. The emulation relying on the decision function of a support vector machine may not correspond to the rating by a real user. The emulation of GRE, COO, OPT and RST may be overly optimistic but MIN can serve as a baseline because it does not employ the decision function.

4 Setting of the Study

4.1 ground truth Databases and Description of the Visual Content

Ground truth image databases are used for evaluating the selection criteria and the user strategies described above; for every database, the ground truth consists in the definition of a set of crisp classes (mutually exclusive here), covering the entire database. For a ground truth database a user can usually find many other classes overlapping those of the ground truth, so the evaluation of a retrieval algorithm 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 databases and to have complexity differences not only among the databases, but also among classes of each database.

Since relevance feedback algorithms must help reducing the semantic gap, the databases should not have too many "trivial" classes, i.e. for which simple low-level visual similarity is sufficient for correct classification. This could happen if the classes were produced for evaluating simple queries by visual example. With these criteria in mind, the first two databases employed here are:

- GT72, composed of the 52 most difficult classes—in terms of internal diversity within classes and of separability between classes—from the Columbia color database, each class containing 72 images.
- GT100, having 9 classes, each composed of 100 images selected from the Corel database. The internal diversity of the classes is stronger than for GT72.

While both GT72 and GT100 are difficult for queries by visual example, every class in these databases can be relatively well modeled by a unimodal distribution. To bring in more complexity, two ground truth databases were designed where each class has several modes:

- GT9F contains 43 classes composed of 2, 3 or 4 sub-classes of 9 images each (see Fig. 1). Every sub-class consists of images selected from several sources and has a visual coherence. Some sub-classes are grouped into classes

according to visual similarity, other according to a more semantic similarity.
- GT30F contains 27 classes composed of 2, 3 or 4 sub-classes of 30 images each (see Fig. 2). As for GT9F, every sub-class consists of images selected from several sources (Web Museum, Corel, Vistex). However, for GT30F there is more internal diversity within each sub-class. The criteria for grouping sub-classes into classes are similar to those employed for GT9F.

The difficulty of the GT9F and GT30F databases comes both from the separation between the different modes of a class and from the presence, in-between some modes, of images belonging to other classes (see Fig. 1). The relevance feedback algorithm must not only succeed in finding, for every class, the other modes that may not be near to the mode of the first "relevant" image, but also be able to exclude "intruders" that belong to other classes; the resulting shape of a class can be rather complex.

The choice of GT9F and GT30F is not only explained by their additional complexity. In real-world retrieval, the starting point of a search session may not belong to the target class of the user, who may have to progressively "guide" the system toward this class, based on a subjective visual similarity. But the crisp nature of the classes found in ground truth databases does not allow to emulate this "focusing" stage of a search session. The presence of several modes in the classes of GT9F and GT30F is then also an attempt to include the constraint of such real-world behavior into the ground truth-based evaluation.

The following signatures are employed for the description of the visual content of the images: a classic HSV color histogram, a Laplacian weighted color histogram [11], a probability weighted color histogram [11], a texture histogram [12] relying on the Fourier transform and a shape feature [12] inspired by the Hough transform. Weighted color histograms are a low-cost solution for taking into account local color uniformity. The texture histogram is based on the application of the Fourier transform to an image and describes the presence of different frequencies along various angles. To obtain the shape feature for a color image, the gray-level image is first computed, then the direction of the gradient is found for every pixel and a reference point is considered; for every pixel, the angle of the gradient and the length of the projection of the reference point along the tangent line going through the pixel position are counted in a joint histogram that is the shape feature.

The complete feature vector is the concatenation of the five types of feature vectors and its dimension is higher than 600, which could make relevance feedback impractical. Linear Principal Component Analysis is a generic unsupervised method for dimension reduction (see e.g. [13]). For the databases and the descriptors considered here, by applying this method and keeping only 95% of the variance in the data, the dimension could be divided by 5.

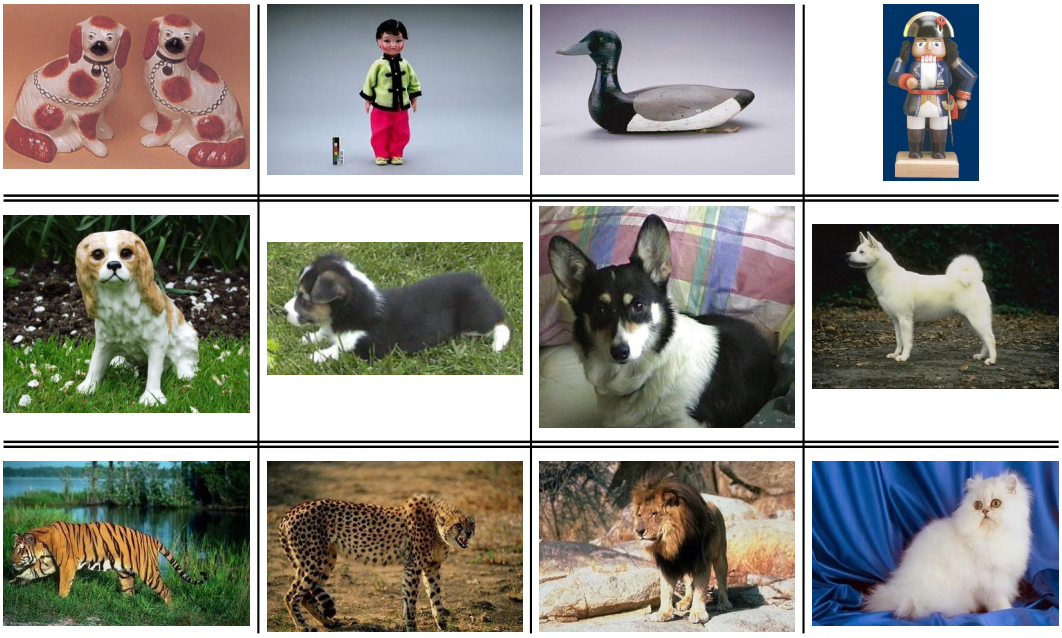

Fig. 1. A sample of images from GT9F. Every line corresponds to a class ("collector", "dog" and respectively "feline") and the different images on a same line come from different sub-classes of that class. The sub-classes of "collector" are better separated than those of the "dog" class. The first sub-class of "collector" may have been added to the "dog" class as well. In the description space, one mode of the "feline" class is between two modes of the "dog" class.

Fig. 2. A sample of images from GT30F. Every line corresponds to a class ("portrait" and respectively "human") and every pair of images on a same line comes from a different sub-class of that class. As compared to GT9F, there is also more diversity within each sub-class.

To allow a reliable comparison of selection criteria and of user strategies, the relevant discriminating information of the ground truth should continue to be represented in the reduced feature vector. It was checked that this strong dimension reduction did not produce a significant loss on the precision/recall diagrams built for queries by visual example.

4.2 Evaluation Method

For all the four databases, at every feedback round the emulated user must label images displayed in a window of size $ws = 9$. Every search session is initialized by considering one "relevant" example and $ws - 1$ "irrelevant" examples. Every image in the database serves as the initial "relevant" example for a different relevance feedback session, while the associated initial $ws - 1$ "irrelevant" examples are randomly selected. In the following evaluations, the focus is 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 ground truth only provides a crisp class membership, the precise ranking of the "relevant" images is ignored.

To evaluate the speed of improvement of this ranking, the precision measure should not give a prior advantage to one selection criterion, nor to some user strategies. Concerning the selection criteria, if precision was defined by counting at every round the already labeled images plus those selected for being labeled during this round, then the MP criterion would be favored over the MAO criterion. Instead, the precision measure is computed as follows: let n be the number of images in the target class; at every feedback round, count the number of images from the target class that are found in the n images considered as most positive by the current decision function of the support vector machine; this number is then divided by n. The "mean precision" reported in all the figures presented here is obtained by averaging the precision measure defined above over all the images in the database, each serving as the initial "relevant" example for a retrieval session.

Regarding the fair comparison of the user strategies defined in Section 3, one can notice that strategies requiring the user to label more images (STO and TIR, followed by ANN, GRE and RST) are favored if the precision measure is computed in terms of iterations (or rounds). Indeed, for any given number of rounds, these strategies provide many more examples to the learner than the other strategies. Computing the precision measure in terms of *clicks* may then seem more equitable. Nevertheless, it can be argued—and this is specific to images—that the time a user needs for evaluating the relevance of all the images in a window is less than proportional to the number of images. Both precision measures are used here: by iterations and by clicks. Since the infor-

mation regarding the precision is only available on a by iteration basis, linear interpolation provides the evolution of precision by click for the user strategies that label more than one image during each round.

Measuring precision as a function of the number of clicks is more relevant for other types of digital content such as texts, music or videos. In all these cases the evaluation of a content item is expensive: the user may have to read a section of text, listen to a fragment of music or watch a video sequence.

5 Evaluation Results

5.1 Comparison Between the Selection Criteria

Comparisons were performed between the MP and MAO selection criteria described in Section 2, on the four ground truth databases and with the eight user strategies. When the strategy of the user changes, it is noticed that:

- For the eight user strategies, the MAO selection criterion performs consistently better than the MP selection criterion. These differences are directly illustrated in Fig. 3 for the ANN strategy with an iteration-based evaluation and in Fig. 4 for the MIN strategy with a click-based evaluation. For the other strategies, the same conclusion can be drawn by comparing Fig. 5 with Fig. 6 and Fig. 7 with Fig. 8.
- The differences in terms of number of clicks between MP and MAO are higher for OPT, MIN, GRE and COO than for STO, ANN and TIR. This advantage of OPT, MIN, GRE and COO can be explained by the fact that for these strategies the selector benefits from more frequent updates of the estimation of the target class by the learner.

These findings apply to all four databases of different complexities. The influence of the complexity of the classes (separability, shape and presence of several modes) can be summarized as follows:

- As expected, performance in terms of number of clicks tends to decrease as the complexity of the database increases. For instance, to achieve a precision of 90%, with the best user strategy among the eight, the number of clicks required is 12 for GT72, 80 for GT100, 250 for GT9F and 200 for GT30F.
- The ranking of the user strategies is relatively stable with respect to changes in database complexity, both in terms of clicks and in terms of iterations, as shown in Tables 1 and 2.
- With MAO, two groups of strategies can be identified: the first consists of COO, GRE and OPT, the second of STO, ANN and RST. As shown in

Tab. 2 or in Figures 7 and 8, the performances within the first group are very similar for each database. This can be partly explained by the fact that small variations in the value of the decision function of the support vector machine do not allow a reliable identification of the most or the least relevant (or irrelevant) image. The disparity between the two groups increases with the complexity of the database.

- Whatever the complexity of the database, the reduction in performance is not catastrophic with 10% of errors in the labels provided by the user. For click-based evaluations, Figures 7 and 8 show a typical example, with the TIR strategy 40% lower than the best strategy with MP after 30 clicks and 30% lower with MAO. For iteration-based evaluations, it is most natural to compare TIR to STO since both strategies label all the images returned; as shown in Figures 5 and 6, the "noise" introduced by the errors for TIR degrades performance by 10%.

The fact that the comparison between MP and MAO is so stable and consistent both with respect to strategy chosen by the user and with respect to the database leads to conclude, with a rather strong confidence, that the MAO selection criterion should be preferred over MP for relevance feedback based on support vector machines with the angular kernel.

Note that the mean precision continuously increases toward 1 when the number of clicks or iterations is incremented, even when images are grouped into classes according to higher level semantics. This is the case for the GT9F and GT30F databases, where each class has multiple (and sometimes well-separated) modes. This result was not obvious *a priori* and is very encouraging with regard to the use of relevance feedback for the reduction of the semantic gap.

Complementary evaluations performed with other kernels and not reported here show that kernels that are highly sensitive to a scale parameter (such as the Gaussian kernel) perform significantly worse on GT9F and GT30F. A tentative explanation is that with such "local" kernels, distant mode easily fall out of sight. In [6] it was argued that since the spatial scales of user-defined classes cannot be known before attempting retrieval, strong variations can be expected for the performance of relevance feedback if kernels such as the Gaussian one are employed. The result regarding multi-modal classes further encourages to prefer kernels that make the learner insensitive to the scale of the data.

Table 1
Ranks of the eight user strategies on the four databases, with the MP and MPO criteria. Ranks are defined by the mean precision after 10 iterations.

Criterion	MP				MAO			
Database	GT72	GT100	GT9F	GT30F	GT72	GT100	GT9F	GT30F
STO	1	1	1	1	1	1	1	1
ANN	3	3	4	4	2	3	4	3
GRE	2	2	2	3	5	5	3	4
COO	8	8	8	8	8	8	8	8
MIN	6	6	6	6	4	7	7	7
OPT	7	7	7	7	6	6	6	6
RST	5	5	5	5	3	4	5	5
TIR	4	4	3	2	7	2	2	2

Table 2
Ranks of the eight user strategies on the four databases, with the MP and MPO criteria. Ranks are defined by the mean precision after 30 clicks. The same rank was given when differences were too small to be considered reliable (these cases are marked by *).

Criterion	MP				MAO			
Database	GT72	GT100	GT9F	GT30F	GT72	GT100	GT9F	GT30F
STO	6	6	6*	6*	5*	5*	5*	5*
ANN	4	4	6*	6*	5*	5*	5*	5*
GRE	3	3	1*	1*	1*	1*	1*	1*
COO	7	7	1*	1*	1*	1*	1*	1*
MIN	2	2	1*	1*	4	4	4	4
OPT	1	1	1*	1*	1*	1*	1*	1*
RST	5	5	5	5	5*	5*	5*	5*
TIR	8	8	8	8	8	8	8	8

The rank between user strategies being surprisingly stable with respect to class complexity and rather similar for the two selection criteria, some user strategies *can* be advised. If the number of clicks is used for evaluation, the best strategies are COO, GRE and OPT, both with MP and MAO.

Increasing the number of negative examples is counter-productive, as shown by the results obtained with RST and STO for click-based evaluations. The classes of images in real generalist databases can have a rather complex shape in the space of image signatures and sometimes several distinct modes, as for GT9F and GT30F. In such cases, relevance feedback can be seen as a process where the user "guides" the system through the description space from already discovered modes to yet undiscovered ones; too many negative examples can block the access to some parts of this space. User strategies that avoid labeling too many negative images tend to perform better in terms of speed of identification of the target class.

With either selection criteria, the GRE strategy appears to be a good trade-off between the number of clicks and the number of iterations. This is also consistent with the above point of view that negative examples *are* necessary but should be employed with care.

6 Conclusion

Relevance feedback is an established method for finding complex, user-defined classes of images. The behavior of real users when labeling images (as "relevant" or not) cannot be expected to follow strict guidelines. An evaluation of the sensitiveness of the retrieval results to likely variations in user behavior was presented here.

Two relevance feedback algorithms based on support vector machines were compared and the user was emulated according to eight significantly different strategies on four ground truth databases of different complexities. It was first found that the ranking of the two algorithms did not depend much on the selected strategy. Second, the ranking between strategies appeared to be relatively independent of the semantic level of the ground truth classes, partly because the kernel employed leads to scale invariance in classification. This robustness to variations in the strategy of the user and in the complexity of the database is a very desirable property when designing systems that should be effective for most users.

Comparisons between relevance feedback algorithms are usually performed using only one user strategy, so it is always questionable whether conclusions extend to real users or not. The comparisons should be conducted with several different strategies—such as the ones put forward above—and the stability of the results evaluated with respect to changes in user strategy and in the complexity of the database. It was also found that user strategies that avoid labeling too many negative examples could be advised to real users because they perform systematically better than the other strategies evaluated.

Finally, in the experiments presented it was noticed how important the choice of the kernel was for relevance feedback. The interaction between the choice of the kernel and user strategies deserves further investigation, in particular for kernels belonging to a larger family of kernels leading to scale invariance.

References

[1] 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*, chapter 8. Prentice Hall, 2004.

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

[3] 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*, pages 750–753, Vancouver, Canada, September 2000.

[4] 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, Ottawa, Canada, 2001. ACM Press.

[5] 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.

[6] Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa. Retrieval of difficult image classes using SVM-based relevance feedback. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 23–30, New York, USA, October 2004.

[7] 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*, Nice, France, October 2003.

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

[9] 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, Stanford, CA, US, 2000. Morgan Kaufmann.

[10] 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, Stanford, CA, US, 2000. Morgan Kaufmann.

[11] 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)*, pages 25–29, Rocquencourt, France, 2001.

[12] Marin Ferecatu. *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*. PhD thesis, Université de Versailles, France, 2005.

[13] Ian T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

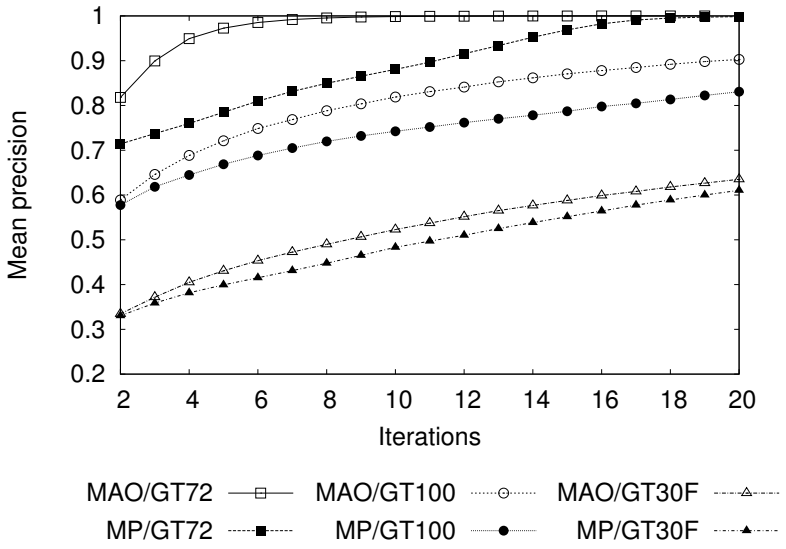

Fig. 3. Iteration-based comparison between the MP and MAO selection criteria with the ANN user strategy on the different databases. The results on GT9F are not shown because they are very similar to those on GT30F.

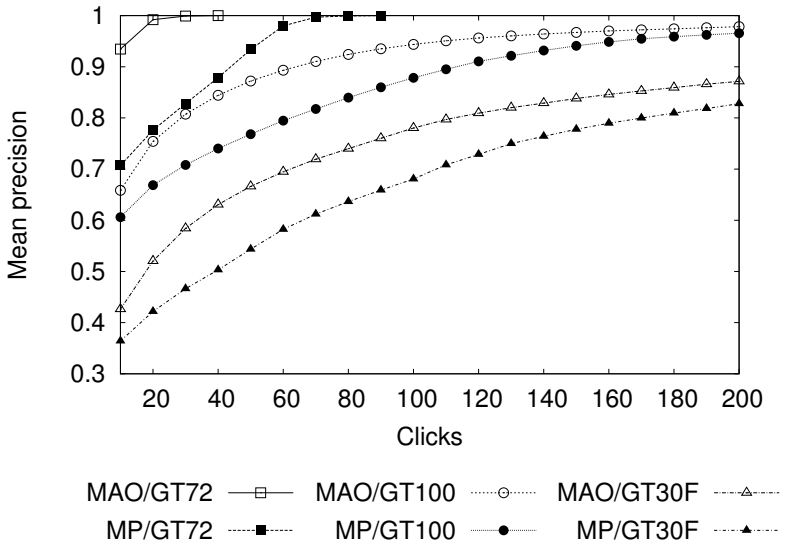

Fig. 4. Click-based comparison between the MP and MAO selection criteria with the MIN user strategy on the different databases. The results on GT9F are not shown because they are very similar to those on GT30F.

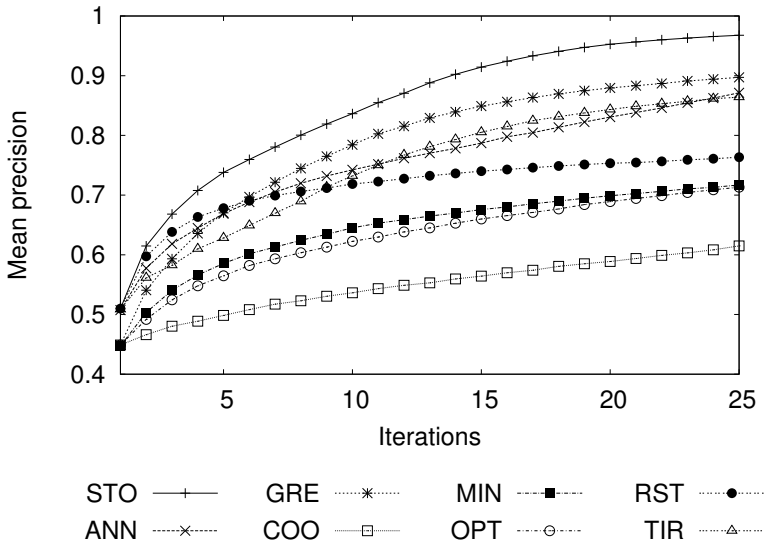

Fig. 5. Iteration-based comparison of the 8 user strategies on the GT100 database, using the MP selection criterion.

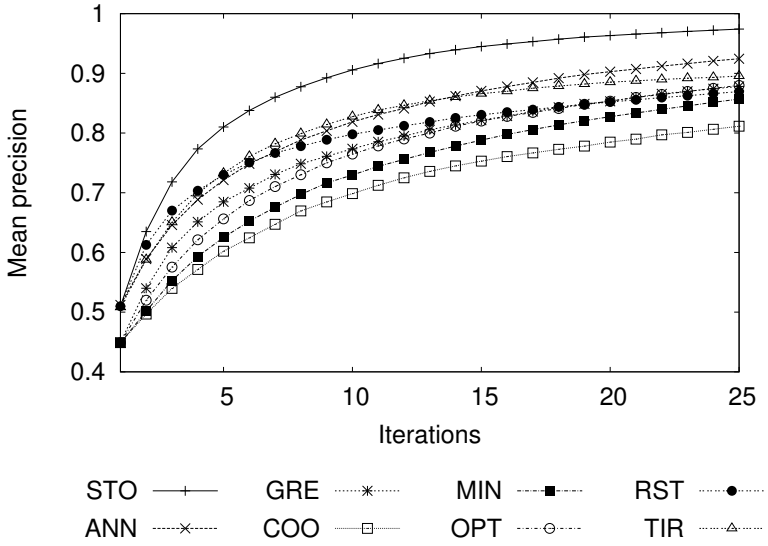

Fig. 6. Iteration-based comparison of the 8 user strategies on the GT100 database, using the MAO selection criterion.

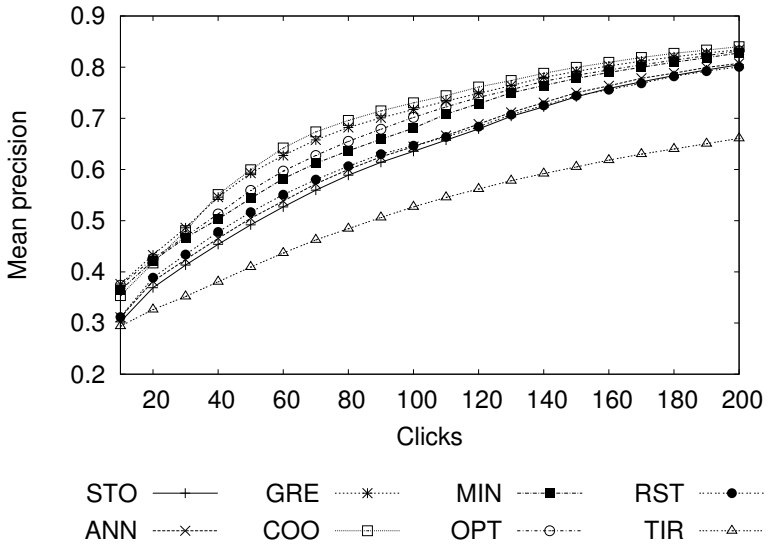

Fig. 7. Click-based comparison of the 8 user strategies on the GT30F database, using the MP selection criterion.

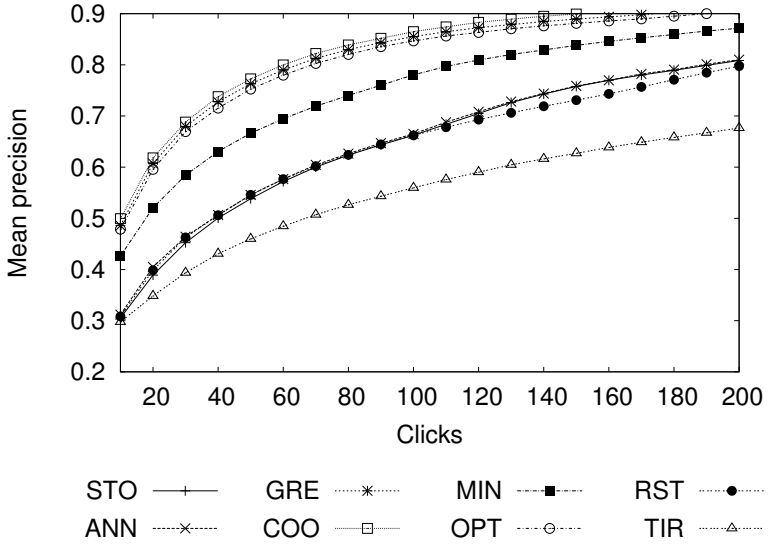

Fig. 8. Click-based comparison of the 8 user strategies on the GT30F database, using the MAO selection criterion.