A Comparison of User Strategies in Image Retrieval with Relevance Feedback

Michel Crucianu, Jean-Philippe Tarel, and Marin Ferecatu

INRIA Rocquencourt, 78153 Le Chesnay cedex, France
{Michel.Crucianu, Jean-Philippe.Tarel, Marin.Ferecatu}@inria.fr

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 an 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. We compare here two algorithms of SVM-based relevance feedback using the angular kernel. In these experiments, the user is emulated according to seven significantly different strategies on four ground-truth databases of different complexity. We first find that the ranking of the two algorithms does not depend much on the selected strategy. Second, the ranking between strategies appears to be relatively independent of the complexity of the ground-truth classes, which allows us to identify desirable characteristics in the behavior of the user.

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 similarity-based search by content and to be able 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* (RF), a session is divided into several consecutive rounds and during every such round 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 RF methods developed, mostly in the content-based image retrieval (CBIR) community, endeavor to minimize the amount of interaction required for ranking most of the "relevant" images before "irrelevant" ones.

Large-scale experiments with real users are costly and difficult to set up, so evaluations and comparisons of RF 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 seems reasonable to expect that the choice of a strategy has an impact on the quality of the RF results. How general, and thus meaningful, are then the conclusions drawn from comparisons performed with ground-truth databases? This is the main issue we study in the following. For cost reasons, we also emulate the user behavior and rely on ground-truth databases. However, in our evaluation, we use multiple strategies as well as several image databases of different complexity.

While such an evaluation cannot replace large-scale experiments with real users, it allows us to explore the impact of various user strategies, at low cost and in a controlled way. We expect this study to bring more confidence to the comparisons between RF algorithms and to provide some insight into possible relations between the RF 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 RF algorithms we compare. The seven user strategies we study and the four ground-truth databases we employ are described in sections 3 and 4, respectively. The results of all these comparative evaluations are presented and discussed in Section 5.

2 SVM-based Relevance Feedback for Image Retrieval

Assume that every image is represented by a signature describing its visual content (see section 4.1). An RF 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" or "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 recent work on RF relies on the use of Support Vector Machines (SVM) [2] to discriminate between "relevant" and "irrelevant" images (e.g. [3–5]). SVMs map the data (image signatures here) to a higher-dimensional feature space (HDFS) using a non-linear transformation associated to a kernel, then implicitly perform linear discrimination between "relevant" and "irrelevant" items in this HDFS. 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 HDFS. This decision function can be used for ranking all the images in the database and deciding which are the most "relevant" ones (with the highest positive values).

Most studies consider the Gaussian (or Radial Basis Function, RBF) kernel, $K(x_i, x_j) = \exp\big(-\gamma\|x_i - x_j\|^2\big)$, with a fixed value for the scale parameter γ (the inverse of the variance). The high sensitivity of the RBF kernel to the scale parameter is an important drawback for RF [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], an interesting alternative is to use the "angular kernel" $K(x_i, x_j) = -\|x_i - x_j\|$. The angular kernel is *conditionally* positive definite, but the convergence of SVM remains guaranteed. In [7], this kernel was shown to have the interesting property of making the frontier found by SVM invariant to the scale of the data (within the limits set by the regularization bound C).

In most RF systems, the selection consists in choosing the images currently considered by the learner to be the most relevant. We call this criterion the selection of the "Most Positive" (MP) candidates. The *active learning* framework for RF using SVMs was introduced in [8, 9]. The associated selection criterion consists in choosing the images whose signatures are the closest to the current frontier between "relevant" and "irrelevant". We call this the selection of the "Most Ambiguous" (MA) candidates. A drawback of MA is that very similar images may be selected. 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 x_i and x_j are the signatures of two candidate images. To have x_i and x_j far apart, a low value for $K(x_i, x_j)$ is required, since the value of the angular kernel decreases with an increase of the distance $d(x_i, x_j)$. We call "MAO" the inclusion of this further condition in the MA criterion.

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 x_j that minimizes the highest value of $K(x_i, x_j)$ for all x_i already included in the current MAO selection: $x_j = \text{argmin}_{x \in S} \max_i K(x, x_i)$. S is the set of images selected by MA and not yet included in the MAO selection, while x_i are the images already in the MAO selection. The number of unlabeled images pre-selected 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). Previous experiments found a value of $2 \times ws$ to be a good compromise and we also employ it here. We compare in the following the MP and MAO selection criteria, using SVM classifiers with the angular kernel.

3 User Strategies

The evaluation and the comparison of RF algorithms usually rely on an emulation of the user following this strategy: given a target class of a ground-truth database, the user qualifies *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; we call this a "stoic" user (STO below).

For cost reasons, we also emulate the users and rely on ground-truth databases, but we investigate, in a controlled way, variations on the behavior of the users, by defining the following six new strategies:

1. An "annoyed" user (ANN) labels only a fixed ratio (50% in the experiments below) of the images returned by the selector; the images it labels are randomly chosen, but the user makes no mistakes when labeling them.
2. A "greedy" user (GRE) correctly labels all the "relevant" images (images belonging to the target class), if present, together with one (randomly chosen) "irrelevant" image (not belonging to the target class), if 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. Since no degree of relevance is available in ground-truth databases, this measure is given here by the SVM decision function: the more positive its value is, the higher the "relevance", and the more negative its value is, the higher the "irrelevance".
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. Here, the most ambiguous "relevant" image is the one for which the value of the decision function is positive but closest to 0.
6. 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 many other variations can be found, we consider that these strategies cover the most important variations expected for the behavior of the users.

We stress that since the user is emulated and our ground-truth databases only contain binary information regarding class membership, we must rely on the decision function of the learner to select the most (or the least) "relevant" (or "irrelevant") among the images returned. This way of evaluating "relevance" is in general not related to the way a real user would rate a "degree of relevance", so the results obtained with COO and OPT should be interpreted with care.

4 Setting of the Study

4.1 Ground-truth Databases and Description of Visual Content

We use ground-truth image databases to evaluate the selection criteria and the user strategies described above; for every database, the ground truth is the definition of a set of binary classes (mutually exclusive here), covering the entire database. Note that 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 paramount to use several databases and to have complexity differences not only among the databases, but also among classes of each database.

Since RF algorithms must help reducing the semantic gap, we try to avoid having in the databases too many "trivial" classes, i.e. for which simple low-level visual similarity is sufficient for correct classification. This is usually the case when the classes are produced for evaluating simple Queries By Visual Example (QBVE). With these criteria in mind, our first two databases 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 has 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 QBVE, every class in these databases can be modeled by a unimodal distribution. To bring in more complexity, we built two ground-truth databases where each class has several modes:

- GT9F contains 43 classes composed of 2, 3 or 4 sub-classes of 9 images each. Every sub-class is composed of images selected from several sources and has a strong 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. As for GT9F, every sub-class is composed of images selected from several sources (Web Museum, Corel, Vistex). However, for GT30F there is more internal diversity within the sub-classes. The criteria for grouping sub-classes into classes are similar to those employed for GT9F.

The difficulty of the GT9F and GT30F databases can be explained both by the separation between the different modes of a class and by the presence of elements from other classes in-between these modes. The RF algorithm must not only succeed in finding the other modes of a class that may not be near to the mode of the first "relevant" image, but also be able to exclude "intruders" from other classes; the resulting shape of a class can be rather complex.

Our 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, so he may have to progressively "guide" the system toward this class, based on his subjective visual similarity. But the binary nature of the classes found in ground-truth databases does not allow for such "focusing" strategy of the emulated user. 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.

For the description of the visual content of the images, we use a Laplacian weighted histogram, a probability weighted histogram, a shape histogram based on the Hough transform, a classic HSV color histogram and a texture histogram based on the Fourier transform. The complete feature vector is the concatenation of individual feature vectors and has more than 600 dimensions, which could make RF impractical. We use a linear PCA to reduce the dimension of the feature vector more than 5 times without a significant loss ($\leq 5\%$) on the precision/recall diagrams in a query by example evaluation.

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 serves as the initial "relevant" example for a different RF session, while the associated initial $ws - 1$ "irrelevant" examples are randomly selected. In our evaluations, we focus 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 only a binary class membership is available, the precise ranking of the "relevant" or of the "irrelevant" images is not important.

To evaluate the speed of improvement of this ranking, we must use a measure that does not give a prior advantage to one selection criterion, nor to some user strategies. Concerning the selection criteria, one can see that if precision is defined by counting at every round the already labeled images plus those selected for being labeled during this round, the MP criterion would be favored over the MAO criterion. We 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 target class).

Regarding the fair comparison of the user strategies defined in section 3, one can notice that strategies requiring the user to label more images 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. We then use both precision measures, in terms of iterations and clicks; since the information regarding the precision is only available on a by iteration basis, for the user strategies that label more than one image during each round we use linear interpolation to obtain the evolution of precision by clicks.

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 by the user is costly: he must read a section of text, listen to a fragment of music or watch a video sequence.

5 Evaluation Results

5.1 Comparison Between Selection Criteria

We performed comparisons between the MP and MAO selection criteria described in section 2, on the four ground-truth databases and with the seven user strategies. When the strategy of the user changes, we noticed that:

- The ranking between MP and MAO, in terms of number of clicks as well as in terms of number of iterations, does not change with the user strategy.

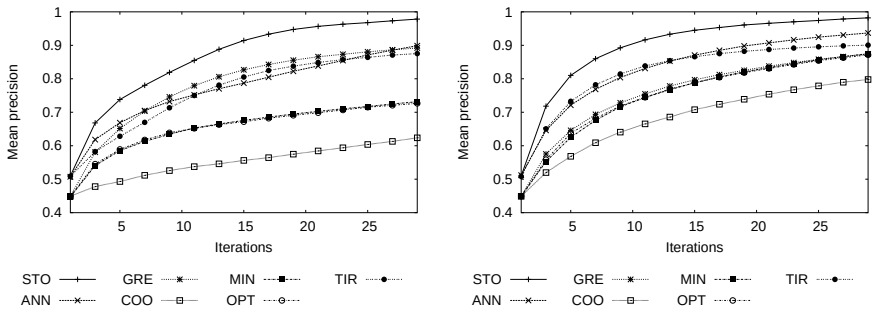

Fig. 1. Evolution of the mean precision with the number of iterations for the 7 user strategies on the GT100 database, using the MP (left) and MAO (right) criteria.

- The MAO selection criterion performs constantly better than the MP selection criterion. Comparison of Fig. 1(a) with Fig. 1(b) illustrates this fact in terms of number of iterations. In terms of number of clicks, the advantage of MAO over MP is even more important, as shown in Fig. 2.
- 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 partially 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 (shape and presence of several modes, separability) can be summarized as follows:

- While MAO performs constantly better than MP, the difference between them appears to increase when the complexity of the classes increases.
- Quite naturally, 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 seven, 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.
- Differences between user strategies are smaller in terms of clicks than in terms of iterations, as illustrated in Tab. 2 and the four figures, whatever the complexity of the database is. Especially using MAO, two groups of strategies can be identified: the first consists in MIN, GRE, OPT and COO, the second in STO and ANN. As shown in Tab. 2 and in Fig. 2, 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 SVM decision function do not allow a reliable decision as to which is the most or the least relevant image. It is also interesting to see that the disparity between the two groups increases with the complexity of the database.

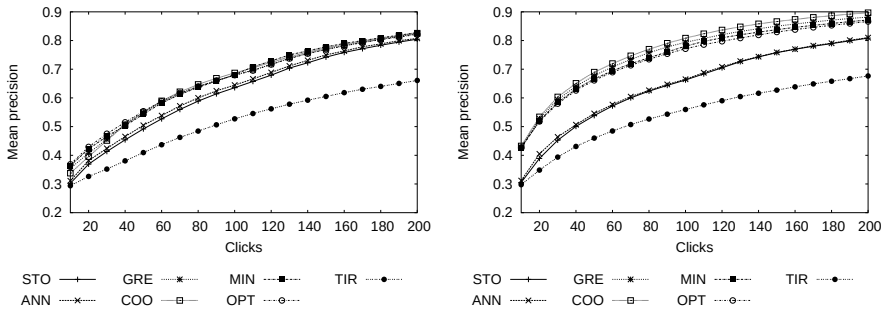

Fig. 2. Evolution of the mean precision with the number of *clicks* for the seven user strategies on the GT30F database, using the MP (left) and MAO (right) criteria.

– Whatever the complexity of the database, the reduction in performance is not catastrophic with 10% of errors in the labels provided by the user. Fig. 2 is a typical example, with the TIR strategy 40% lower than the best strategy with MP after 30 clicks and only 30% lower with MAO.

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 us to conclude, with a rather strong confidence, that the MAO selection criterion should be preferred over MP for SVM-based RF with the angular kernel. It is also important to notice that the average precision always converges toward 100% when the number of clicks or iterations increases, even when images are grouped into classes according to higher level semantics. This is often the case for the GT9F and GT30F databases, where each class has multiple modes. This result was not obvious *a priori* and is very encouraging with regard to the use of RF for the reduction of the semantic gap.

Table 1. Ranks of the seven user strategies on the four databases, with the MP and with the MPO criterion. 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	3	4	3	4
COO	7	7	7	7	7	7	7	7
MIN	6	6	6	6	5	6	6	6
OPT	5	5	5	5	4	5	5	5
TIR	4	4	3	2	6	2	2	2

Table 2. Ranks of the seven user strategies on the four databases, with the MP and with the MPO criterion. Ranks are defined by the mean precision after 30 clicks. When differences are too small to be reliable, a same rank is given.

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

5.2 Advisable User Strategies

The rank between user strategies being surprisingly stable with respect to class complexity and rather similar with the two selection criteria, we consider that some user strategies *can* be advised. In the number of clicks is used for evaluation, with the MP criterion, the best strategies are OPT and MIN, while with the MAO criterion, GRE and COO can also be included.

User strategies that maintain a balance between positive and negative examples are not the best-performing ones. Increasing the number of negative examples appears to be counter-productive. 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. Then, RF can be seen as a process where the user "guides" the system through the description space and 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 convergence of the RF toward the target class.

When using the best-performing MAO selection criterion, 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. This suggests that the use of the GRE strategy (see section 3) should be advised to real users.

6 Conclusion

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

We compared two algorithms of SVM-based relevance feedback and we emulated the user according to seven significantly different strategies on four ground-

truth databases of different complexities. We first find that the ranking of the two algorithms does not depend much on the selected strategy. Second, the ranking between strategies appears to be relatively independent of the semantic level of the ground-truth classes, thanks in part to the choice of a kernel that 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. We suggest that the comparisons should be conducted with several different strategies—such as the ones we put forward here—and the stability of the results evaluated with respect to changes in user strategy.

We also find that user strategies that avoid labelling too many negative examples perform systematically better than the other strategies we evaluated. The GRE strategy could thus be advised to real users.

Finally, in our experiments we noticed how important the choice of the kernel was in SVM-based relevance feedback. In future research we will focus on the interaction between the choice of the kernel and user strategies, in particular for kernels belonging to a larger family of kernels leading to invariance to scale.

References

1. Gevers, T., Smeulders, A.W.M.: Content-based image retrieval: An overview. In Medioni, G., Kang, S.B., eds.: Emerging Topics in Computer Vision, Prentice Hall (2004)
2. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press (2002)
3. Hong, P., Tian, Q., Huang, T.S.: Incorporate support vector machines to content-based image retrieval with relevant feedback. In: Proceedings of the 7th IEEE International Conference on Image Processing. (2000)
4. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: Proceedings of the 9th ACM International Conference on Multimedia, ACM Press (2001) 107–118
5. Jing, F., Li, M., Zhang, H.J., Zhang, B.: Learning region weighting from relevance feedback in image retrieval. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. (2002)
6. Ferecatu, M., Crucianu, M., Boujemaa, N.: Retrieval of difficult image classes using SVM-based relevance feedback. In: Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval. (2004) 23–30
7. Fleuret, F., Sahbi, H.: Scale-invariance of support vector machines based on the triangular kernel. In: 3rd International Workshop on Statistical and Computational Theories of Vision. (2003)
8. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: Proceedings of ICML-00, 17th International Conference on Machine Learning, Morgan Kaufmann (2000) 999–1006
9. Campbell, C., Cristianini, N., Smola, A.: Query learning with large margin classifiers. In: Proceedings of ICML-00, 17th International Conference on Machine Learning, Morgan Kaufmann (2000) 111–118