Sample selection strategies for relevance feedback in region-based image retrieval

Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa

INRIA Rocquencourt, 78153 Le Chesnay Cedex, France
Marin.Ferecatu@inria.fr, Michel.Crucianu@inria.fr,
Nozha.Boujemaa@inria.fr

Abstract. The success of the relevance feedback search paradigm in image retrieval is influenced by the selection strategy employed by the system to choose the images presented to the user for providing feedback. Indeed, this strategy has a strong effect on the transfer of information between the user and the system. Using SVMs, we put forward a new active learning selection strategy that minimizes redundancy between the examples. We focus on region-based image retrieval and we expect our approach to produce better results than existing selection strategies. Experimental evidence in the context of generalist image databases confirms the efectiveness of this selection strategy.

1 Introduction

The concept of *semantic gap* has been extensively used in the Content Based Image Retrieval (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 to the users of the search engines [1].

Image regions are usually perceived as beeing closer to semantic concepts and one way to address the semantic gap is to concentrate the search at the region level where a relation between concepts and regions is easier to establish [2]. Another solution for reducing the semantic gap is to cut a search session into several consecutive retrieval rounds (iterations) and let the user provide feedback regarding the results of every retrieval round, e.g. by qualifying images returned as either "relevant" or "irrelevant" [3] (relevance feedback, RF).

In order to maximize the ratio between the quality (or relevance) of the results and the amount of interaction between the user and the system, the selection of images for which the user is asked to provide feedback at the next round must be carefully studied. For a *target search* scenario, where the user is searching for a specific image, interesting ideas were introduced in [4]. At every round, the user is required to choose between two images presented by the engine and the selection strategy must let the user remove a maximal amount of uncertainty regarding the target. We consider that this criterion translates into two complementary conditions for the images in the selection: (1) each image must be ambiguous given the current estimation of the target and (2) the redundancy between the different images has to be low. However, computational optimizations are required for searching larger sets of images (*category* search) and for selecting more than 2 images.

Based on the definition of active learning (see for example [5]), the selection of examples for training SVMs to perform general classification tasks is studied in [6]. In the early stages of learning, the classification of new examples is likely to be wrong, so the fastest reduction in generalization error can be achieved by selecting the example that is farthest from the current estimation of the frontier. During late stages of learning, the classification of new examples is likely to be right but the margin may be suboptimal, so the fastest reduction in error can be achieved by selecting the example that is closest to the current estimation of the frontier. Note that, according to the classical formulation of active learning, the authors only consider the selection of single examples for labeling (or addition to the training set) at every round.

Also for SVM learners, several selection criteria are presented in [7] and applied to content-based text retrieval with relevance feedback. The simplest (and computationally cheapest) of these criteria consists in selecting the texts whose representations (in the feature space induced by the kernel) are closest to the hyperplane currently defined by the SVM. We shall call this simple criterion the selection of the "most ambiguous" (MA) candidate(s). This selection criterion is justified in [7] by the fact that knowledge of the label of such a candidate halves the version-space (the set of learner parameters that are compatible with the already labeled examples). In order to minimize the number of feedback rounds, the user is asked to label several examples at every round and all these examples are selected according to the MA criterion. In [8] 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 further labeling.

In the next section we put forward a new active selection strategy based on the reduction of the redundancy between the examples presented to the user. Experimental evidence in Section 3 shows that our strategy performs well compared with other strategies in generalist database region based query contexts. Concluding remarks are given in Section 4.

2 Reduction of the Redundancy

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 candidates.

We suggest here to translate this condition of low redundancy into the following additional condition: if x_i and x_j are the input space representations of two candidate images, then we require a low value for $K(x_i, x_j)$ (i.e. of the value taken by the kernel for this pair of images). If the kernel K is inducing a Hilbert structure on the feature space, if $\phi(x_i)$, $\phi(x_j)$ are the images of x_i,

x_j in this feature space and if all the images of vectors in the input space have constant norm, then this additional condition corresponds to a requirement of quasi-orthogonality between $\phi(x_i)$ and $\phi(x_j)$ (since $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$). We shall call this criterion the selection of the "most ambiguous and orthogonal" (MAO) candidates.

We note that the MAO criterion can be extended to reduce redundancies between the examples selected during subsequent RF rounds. This additional constraint may be important in situations where the number of labeled examples is much lower than the dimension of the input space and the classes are restricted in most directions.

The MAO criterion has a simple intuitive explanation for kernels $K(x_i, x_j)$ that decrease with an increase of the distance $d(x_i, x_j)$ (which is the case for most common kernels): it encourages the selection of unlabeled examples that are far from each other in input space, allowing to better explore the current frontier.

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

$$x_j \ = \ \mathrm{argmin}_{x \in S} \ \mathrm{max}_i \ K(x, x_i)$$

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

In a general classification context, a similar "diversity" condition for the selected examples was put forward in [9] and evaluated on several benchmark classification problems from the UCI database. The condition is justified by reference to the version space account suggested in [7]: diversity is maximized when the hyperplanes associated to the individual examples are orthogonal and are thus complementary to each other in halving the version space.

We note that the MA criterion in [7], [8] is the same as the one put forward in [6] 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.

In early stages of the learning the frontier is very unreliable and selecting those unlabeled examples that are currently considered by the learner as (potentially) the most relevant can sometimes produce a faster convergence of the frontier during the first few rounds of RF.

For this reason, we added to our comparisons the following criteria: select the "most positive" unlabeled examples according to the current decision function of the SVM (denoted as MP criterion) and select the "most positive and orthogonal" unlabeled examples (denoted as MPO). The MPO criterion adds to MP the condition of low redundancy previously described. When comparing the MP criterion to the suggestion in [6] for the early stages of learning, we see that we only focus on the examples for which the values taken by the decision function of the SVM are maximal and completely ignore the examples for which these values are minimal; this is because of the asymmetry of the retrieval context: in general, the number of relevant items is expected to be much lower than the number of irrelevant items.

3 Experimental Evaluation

To test our selection criterion we selected a groundtruth database of image regions, built from a generalist image database. The first stage was to automatically obtain a coarse segmentation of the images in the database using the algorithms described in [2].

To describe the visual content of image regions we used Laplacian wighted histograms, probability weighted histograms, shape histograms based on the Hough transform and classic HSV color histograms. Weighted color histograms rely on the idea that not all pixels are equal when it comes to their contribution to the histogram. Pixels from uniform regions of an image are less important than pixels from regions where there are important changes in color (see [10]). These histograms integrate a local measure of the uniformity of the pixels, and thus have a texture description value added to their primary intended color description.

The final feature vector is the concatenation of individual feature vectors and has more than 600 dimensions. The very high number of dimensions of the joint feature vector can make RF impractical even for medium-size databases. Also, the higher the dimensionality of the description space, the more difficult is the task of the learner. We use a linear PCA to reduce the dimension of the feature vector more than 5 times, without a significant loss (less than 5%) on the precision/recall diagrams in a query by example context.

To build the groundtruth, we annotated by hand 5401 regions from a total of 44286 automatically segmented regions in our database. We obtained 27 classes (such as hair, face, sea, forrest, village, shutters, etc.), most of them containing more than 150 regions. The results we show here correspond to the "sea" class (510 image regions) and to the "face" class (523 image regions).

At every feedback round the (emulated) user must label as "relevant" or "irrelevant" all the images in a window of size $ws = 9$. A search session is initialized by considering one "relevant" example and $ws - 1$ "irrelevant" examples. Every image in every class serves as the initial "relevant" example for a different RF session, while the associated initial $ws - 1$ "irrelevant" examples are randomly selected.

For the SVM we employed the triangular kernel, $K(x_i, x_j) = -\|x_i - x_j\|$, because in all our experiments it performed better than other kernels (RBF, Laplace). Also, this kernel has the property of making the frontier found by SVMs invariant to the scale of the data (see [11]). Classes of image regions in generalist databases usually have very different scales in the space of low-level

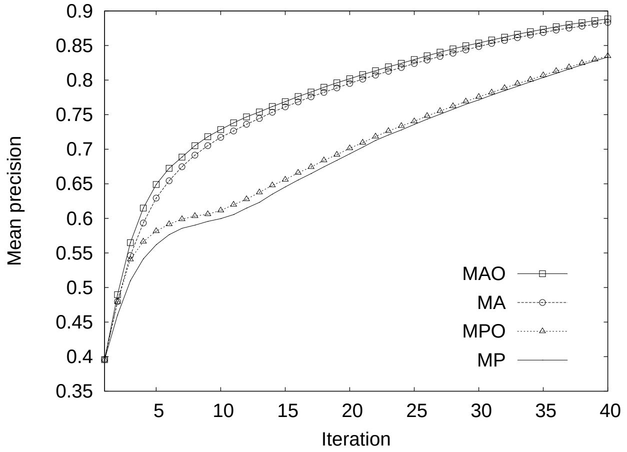

Fig. 1. Comparison of the selection strategies for the "sea" class.

descriptors; thus, kernels producing scale invariance are to be preferred to the classical ones that are very sensitive to a scale parameter.

We first evaluated the different selection criteria in a **ranking scenario**: finding items in a specific target set, by focusing on ranking most of the "relevant" images before the "irrelevant" ones rather than on finding a frontier between the class of interest and the other images. In order to evaluate the speed of improvement of this ranking, we must use a measure that does not give a prior advantage to one selection criterion. For example, by taking into account already labeled images plus those selected for being labeled during the current round, we should obviously favor the MP and MPO criteria over MA and MAO. We decided to use instead the following precision measure: at every RF round, we count the number of "relevant" images found in the first N images considered as most positive by the current decision function of the SVM (N being the number of images in each class).

The evolution of the mean precision during successive RF iterations (rounds) is presented in Fig. 1 for the "sea" class and in Fig. 2 for the "face" class. The "mean precision" value shown is obtained as the mean over all the image regions in the class of the precision measure described above. Clearly, the reduction of the redundancy between the images selected for labeling improves the results, both for MAO with respect to MA and for MPO with respect to MP. Also, in these cases the MA and MAO selection criteria compare favorably to the MP and MPO criteria.

The **second type of scenario** we evaluated consists in finding a frontier between "relevant" and "irrelevant" images, which can be important for extending textual annotations of some images in the "relevant" class to the others. In this case, we have to evaluate the speed of improvement of the classification. The classification error is defined here as $n/N + (N-p)/N$, where N is the class size,

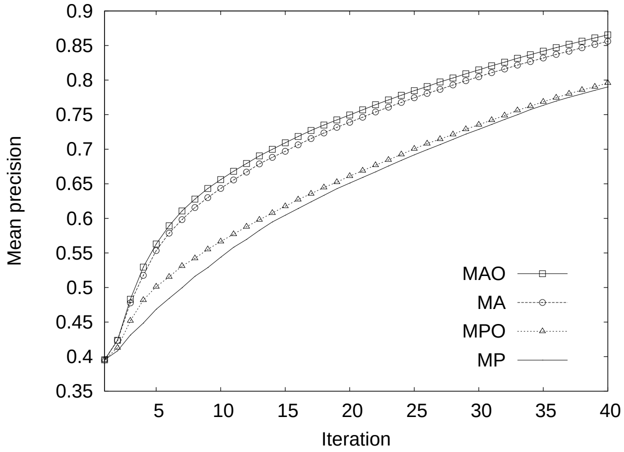

Fig. 2. Comparison of the selection strategies for the "face" class.

n is the number of false positives and p is the number of true positives (thus $N - p$ is the number of false negatives). In Fig. 3 we can see the evolution of the classification error obtained with the different selection criteria for the "sea" class. As expected, the convergence is fastest for the MAO selection criterion, followed by the MA criterion.

Similar results are obtained for most of the other classes, with one remark: simple classes, with a small number of elements concentrated in a relatively small region in feature space, have an almost identical behavior for all the selection strategies. Also, we found only two classes (out of the 27) where the MP criteria performed slightly better that the MA ones: "building" and "city". Principal component analysis reveals that the projections of the two classes are rather spread and biased discriminant analysis shows that these classes are very mixed, making them very difficult to separate by the SVM. This suggests that complementary image features should be used to discriminate the two classes.

4 Conclusion

In content-based image retrieval with relevance feedback, the strategy employed by the search engine for selecting the images presented to the user at every feedback round is very important for the transfer of information between the user and the system. Using SVMs as learners, we put forward an improved active learning selection strategy, based on a reduction of the redundancy between the images selected at every feedback round.

By comparing this strategy to alternative strategies for the retrival of regions in the context of generalist image databaseses, we have shown that it performs better in ranking most of the "relevant" images before the others and also speeds up the convergence of the frontier around the class of interest. This last aspect

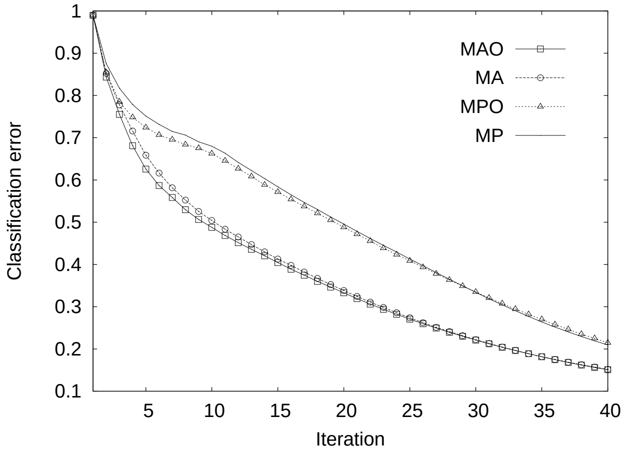

Fig. 3. Evolution of the classification error obtained with different selection strategies for the "sea" class.

is important when relevance feedback is used as a tool for extending semantic annotations.

As a visual example of retrieval, in Fig 4 we present the third screen of results returned by our CBIR system IKONA (see [10]), on the database used in this paper. In this example the class the user is searching for is "village" (having 87 examples) and at this point of the RF session there are 4 positive examples and 9 negative examples already annotated.

In this case, all the regions returned belong to the "village" class; nevertheless they have different characteristics: some are close-ups and some are not, the texture is very different, some focus on single buildings and some are global views, etc. Moreover, this class is easy to confuse with other classes ("rock", "mountain", "city") in terms of color, texture and shapes, characteristics used by the image features.

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. Fauqueur, J., Boujemaa, N.: Region-based image retrieval: Fast coarse segmentation and fine color description. Journal of Visual Languages and Computing (JVLC), special issue on Visual Information Systems **15** (2004) 69–95
3. Zhou, X.S., Huang, T.S.: Relevance feedback for image retrieval: a comprehensive review. Multimedia Systems **8** (2003) 536–544
4. Cox, I.J., Miller, M.L., Minka, T.P., Papathomas, T., Yianilos, P.N.: The Bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments. IEEE Transactions on Image Processing **9** (2000) 20–37

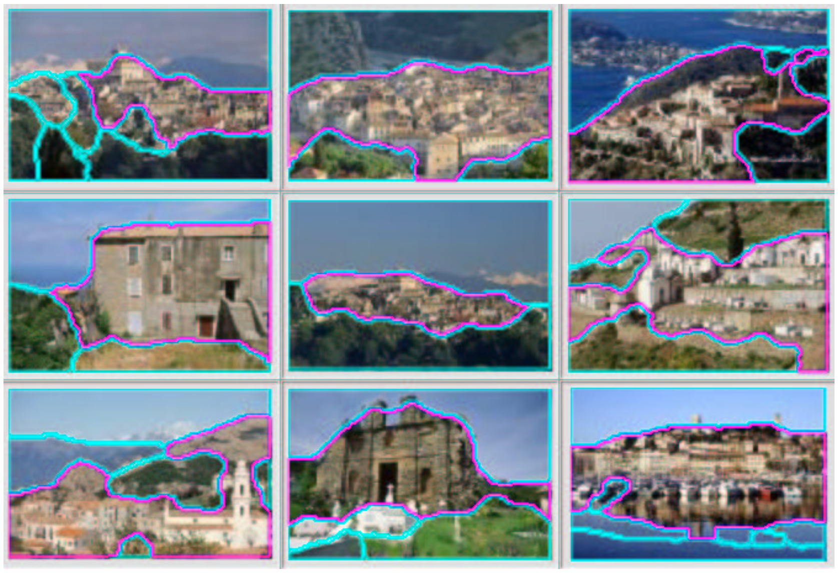

Fig. 4. Searching for village regions in a generalist database. The contours of the regions matching the query are in red (dark gray) and those of the other regions in light blue (light gray).

5. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of Artificial Intelligence Research **4** (1996) 129–145
6. 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
7. 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
8. 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
9. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: Proceedings of ICML-04, International Conference on Machine Learning. (2003) 59–66
10. Boujemaa, N., Fauqueur, J., Ferecatu, M., Fleuret, F., Gouet, V., Saux, B.L., Sahbi, H.: Ikona: Interactive generic and specific image retrieval. In: Proceedings of the International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001). (2001)
11. 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)