SEMANTIC KERNEL LEARNING FOR INTERACTIVE IMAGE RETRIEVAL

Philippe H. Gosselin and Matthieu Cord

ETIS / CNRS UMR 8051
6, avenue du Ponceau, 95014 Cergy-Pontoise, France

ABSTRACT

Content-based image retrieval systems still have difficulties to bridge the semantic gap between the low-level representation of images and the high level concepts the user is looking for. Relevance feedback methods deal with this problem using labels provided by users, but only during the current retrieval session. In this paper, we introduce a semantic learning method to manage user labels in CBIR applications. Our approach uses a kernel matrix to represent semantic information in a statistical learning framework. The kernel matrix is updated according to labels provided by users after retrieval sessions. Experiments have been carried out on a large generalist database in order to validate our approach.

1. INTRODUCTION

Traditional techniques in content-based image retrieval (CBIR) are limited by the semantic gap between the low-level representations of images and the semantic subsets of the database the users are looking for [1, 2]. The user is looking for one image or an image set with semantics, for instance a type of landscape, whereas current processing deals with color or texture features. The increasing database sizes and the diversity of search types contribute to increase the semantic gap.

Recently, statistical learning approaches have been introduced in CBIR context [3] and have been very successful to improve the effectiveness of visual information retrieval tasks. Discrimination methods approach the relevance feedback as a supervised learning problem. A binary classifier is learned by using all relevant and irrelevant labeled images as input training data [4]. The user interactively provides labels, and the classification function is updated. We have introduced an active learning strategy [5] to carry out an efficient relevance feedback working as well with SVM as other classification methods.

Most of the time, the labels provided during one retrieval experiment are not re-used thereafter. In this paper, we introduce a learning method to exploit the information accumulated during past sessions. Our technique, called RETIN SL (Semantic Learning), is working on generalist image databases, and is designed to model mixed categories. In order to learn any kind of semantic links, we use a similarity matrix framework. To keep our general framework easy to combine with other techniques (relevance feedback and active learning, clustering, ...), we use a kernel matrix as a similarity matrix. We introduce some receivable algebraic transformations in order to always keep the nice properties of the kernel. The purpose is to reinforce similarity matrix values between images identically labeled by the user. To handle huge databases, we also propose an approximation technique with a complexity linear to the size of the database.

In this scope, we first present in Section 2 the active learning strategy for relevance feedback. In Section 3, we describe the RETIN SL method of semantic learning. In Section 4, experiments combining RETIN SL and active on-line strategies are reported.

2. ACTIVE LEARNING

Performances of inductive classification depend on the training data set. In interactive CBIR, all the images labeled during the retrieval session are added to the training set used for classification. As a result, the choice of which labeled images to add will change system performances. For instance, labelling an image which is very close to one already labeled will not change the current classification.

Usual strategies in statistical learning propose to choose elements with the less classification accuracy. Some researchers as Cohn [6], propose to train several classifiers with the same training data, and choose data where classifiers disagree at most. Tong adapts SVM framework to active learning strategy in order to get the SVM$_{active}$ learning method [4]. SVM$_{active}$ tries to focus the user on images whose classification is difficult. It asks the user to label m images closest to the SVM boundary ($m = 20$ in their experiments [4]).

Let $(\mathbf{x}_i)_{i\in[1,n]}$, $\mathbf{x}_i \in \mathbb{R}^p$ be the feature vectors representing images from the whole database, and $\mathbf{x}_{(i)}$ the permuted vectors after a sort according to a decision function f. At the feedback iteration t, SVM$_{active}$ proposes to label m images from rank s_t to $s_t + m - 1$ so that $\mathbf{x}_{(s_t)}, ..., \mathbf{x}_{(s_t+m-1)}$ are the closest images to the SVM boundary. SVM$_{active}$ strategy rests on a strong theoretic foundation and increases performances, but it works with an important assumption: a reliable estimation of the boundary between classes.

We introduced a method with the same principle than SVM$_{active}$ but without using the SVM boundary to find the value s_t [5]. Indeed, we notice that, even if the boundary may change a lot during the first iterations, the ranking operation is quite stable. Actually, we just suppose that the best s_t (corresponding to the searched boundary) allows to present as many relevant images as irrelevant ones. Thus, if and only if the set of the selected images is well balanced (between relevant and irrelevant images), then s_t is relevant. We exploit this property to adapt s_t during the feedback steps. Comparisons in [5] have shown the efficiency of the method, especially with few training data.

3. SEMANTIC KERNEL LEARNING

Let us note *semantics* all the information (users' annotations) accumulated from many retrieval sessions. Different strategies may

be used to learn information about the database from these *semantics*:

- Some approaches deal with feature selection or competition [7]. The Latent Semantic Index and its kernel version have also been proposed to model the correlation between feature variables [8].

- Other approaches compute and store a similarity matrix. A lot of approaches are based on the Kernel Alignment [9]. The idea is to adapt a kernel matrix (which is a particular similarity matrix) considering user labelling. This problem can be solved using semi-definite quadratic programming[1] [10]. However, they have been designed mostly for transduction and clustering, $i.e.$, two class problems. For generalist database searches, there are many concepts or categories, overlapping each other. Some methods, building and updating a similarity matrix, have been experimented [11]. Usually, there is no assumption about the properties of the similarity matrix. For instance, the updated matrix may lost the induced metric properties. Moreover, these similarity matrix-based approaches have also a high computational cost. The memory complexity is at least $O(N^2)$, where N is the number of pictures in the database.

Our strategy is based on a kernel matrix adaptation, and is designed to model mixed categories. We also manage the complexity constraint using efficient eigenvalue matrix decomposition; the method has a $O(N)$ complexity and memory need, and so it is applicable to large databases.

3.1. Adaptive approach

Let us note K_t the kernel matrix at the end of the retrieval session t: $(K_{ij})_t = k((\mathbf{x}_i)_t, (\mathbf{x}_j)_t)$. When a kernel function $k(.,.)$ is used, the matrix K_t is symmetric and semi-definite positive (sdp), it is a Gram matrix. We propose algebraic transformations always keeping the sdp property of the kernel matrix.

The labels provided at step t are stored in a vector $\mathbf{y}_t$ of size N, with 1 for relevant pictures, -1 for irrelevant pictures, and 0 for unlabeled pictures. If the system is used several times by users, a set of labels becomes available:

	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	...
$\mathbf{x}_1$	1	1	0	0	-1	1	0	...
$\mathbf{x}_2$	1	1	1	1	-1	0	1	...
$\mathbf{x}_3$	1	0	1	-1	0	0	0	...
$\mathbf{x}_4$	0	-1	1	0	0	-1	0	...
$\mathbf{x}_5$	-1	0	0	1	1	-1	0	...
$\mathbf{x}_6$	0	0	-1	0	1	0	-1	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Labels give a partial information about the category the user has in mind, a large majority of pictures is unlabeled for a given $\mathbf{y}_t$.

After each retrieval process t, the current kernel matrix K_t is updated using the following expression:

$$K_{t+1} = (1-\rho)K_t + \rho K_{\mathbf{y}_t} \tag{1}$$

where the weight $\rho \in [0,1]$ is tuned according to the confidence we have in labels provided by user interaction, and $K_{\mathbf{y}_t}$ a matrix containing information from $\mathbf{y}_t$. This matrix must be sdp so that K_{t+1} keeps the sdp property.

[1] Semi-definite programming allows efficient algorithms.

3.2. Merging K_t and $\mathbf{y}_t$

Two types of operations are considered for $K_{\mathbf{y}_t}$. The first one is the $N \times N$ matrix $K_{\mathbf{u}_t}$, defined as:

$$K_{\mathbf{u}_t} = \mathbf{u}_t(\mathbf{u}_t)' \text{ with } u_{ti} = \begin{cases} 1 & \text{if } y_{ti} > 0 \\ -\gamma & \text{if } y_{ti} < 0 \\ 0 & \text{otherwise} \end{cases}$$

which increases the similarity between positive labeled images, and decreases the similarity between negative and positive labeled images. The $\gamma \in [0,1]$ parameter deals with the increasing of similarity between negative labeled images[2]. $K_{\mathbf{u}_t}$ is a sdp matrix because it is a rank one matrix with a positive eigenvalue ($||\mathbf{u}_t||^2$).

The second one aims at averaging all the similarities between the positive labeled images, with the operator $K_T = TK_tT'$, $N \times N$ matrix T defined as:

$$T = \begin{pmatrix} \frac{1}{q_+} & \cdots & \frac{1}{q_+} & & & \\ \vdots & & \vdots & & & \\ \frac{1}{q_+} & \cdots & \frac{1}{q_+} & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}$$

To simplify the notations, we suppose that the q_+ first values of $\mathbf{y}_t$ are positives, and the q_- next values of $\mathbf{y}_t$ are negatives)

It is also easy to prove the sdp property of K_T, as soon as K_t is sdp, using the definition: M is $sdp \iff \forall \mathbf{x} \in \mathbb{R}^N$, $\mathbf{x}'M\mathbf{x} \geq 0$.

We then propose to merge knowledge in K_t and $\mathbf{y}_t$ using the following operator:

$$K_{\mathbf{y}_t} = a \times (TK_tT' + bK_{\mathbf{u}_t}) \tag{2}$$

with $b \in \mathbb{R}^+$ such as diagonals of $TK_tT' + bK^{\mathbf{u}_t}$ are 1, and $a \in \mathbb{R}^+$ such as $\sum_{ij}(K_{ij})_{t+1} = \sum_{ij}(K_{ij})_t$.

From eq. (1) and (2), the final expression of the kernel updating is the following one:

$$K_{t+1} = (1-\rho)K_t + \rho a(TK_tT' + bK_{\mathbf{u}_t}) \tag{3}$$

3.3. Semantic kernel computation

We use a low-rank approximation of the matrix K_t, in order to have a storage linear to the size of the database. As the kernel matrix is real and symmetric, we are able to compute its eigendecomposition. The approximation consists in keeping the m largest eigenvalues. Thus, assuming that $m \ll N$, the storage of K_t is $O(N)$.

We use a factorization technique for the computation of the eigenspectrum of K_{t+1}. The factorization is followed by a QR decomposition and the computation of the eigenspectrum of a very small matrix (in comparison to N). This method has a $O(N)$ complexity.

[2] In a multiple category context, negative labeled images are usually not is the same category. Thus in this case a small value (0.1) of γ is preferable.

4. EXPERIMENTS

4.1. Protocol and parameters

Tests are carried out on the generalist COREL photo database, which contains more than $50,000$ pictures. To get tractable computation for the statistical evaluation, we randomly selected 77 of the COREL folders, to obtain a database of $6,000$ images. To perform interesting evaluation, we built from this database 11 categories [3] of different sizes and complexities like birds (219), castles (191), doors (199), Europe (627) , food (315), mountains (265) ...

The CBIR system performances are measured using precision(P), recall(R) and statistics computed on P and R for each category. We use the mean average precision (MAP) which represents the value of the P/R integral function. This metric is used in the TREC VIDEO conference[4], and gives a global evaluation of the system (over all the (P,R) values).

The semantic kernel matrix is initialized with the $L^*a^*b^*$ and Gabor features: $K_{t=0} = X'X$, with $X = (\mathbf{x}_i)_{i\in[1,N]}$ the $p \times N$ distribution matrix, for which each column $\mathbf{x}_i$ is a vector representation of the ith picture of the database.

In experiments, we consider 3 active learning strategies: a basic active learner Basic AL, which chooses the best ranked unlabeled pictures, RETIN AL [5], and SVM$_{active}$ [4]. For each active learner, 5,000 retrieval sessions are simulated: one category is randomly selected, the search starts with one random picture from this category, 5 pictures selected by the active learner are automatically labeled, and this process is repeated during 10 feedback steps. At the end of such a retrieval session, the semantic kernel is updated using the semantic learning method with the 50 labels. Every 1,000 retrieval sessions, the semantic kernel K_t is stored.

In simulation, we have 6 kernels for each active learning method: $K^0, K^{1000}, ..., K^{5000}$. For each kernel, system performances are evaluated with the mean average precision for each category. Finally, mean, minimum and maximum MAP for all categories are computed. Results are displayed in Fig. 2, 3 and 4.

The ρ parameter. The method has been evaluated with ρ values 0.01, 0.05, 0.1, 0.5 and 1. As a rule, when ρ increases, the system learns faster. However, over 0.5 the learning becomes unstable: the MAP may increase a lot for some categories, whereas it decreases for other ones.

The γ parameter. The method has been evaluated with γ values 0.01, 0.05, 0.1, 0.5 and 1. The system has the best learning performances when $\gamma = 0.1$. Below this value, the system learns slowly, and above learning is inefficient: with a value of 1, the MAP is decreased.

The number m of non-zero eigenvalues. The method has been evaluated with m values 10, 25, 50, 100 and 200. Globally, the higher it is, the best performances are. However, starting from a given value (here 50), performances do not increase a lot. Furthermore, it seems that the number of eigenvalues is mainly linked to the number of categories users are looking for, not the number of images in database. We experimented the system with 5 categories, and in this case 25 eigenvalues were enough.

In the following experiments, $\rho = 0.1$, $\gamma = 0.1$, and $m = 50$.

[3] A description of this database and the 11 categories can be found at: http://www-etis.ensea.fr/∼cord/data/mcorel.tar.gz. This archive contains lists of image file names for all the categories.

[4] http://www-nlpir.nist.gov/projects/trecvid/

4.2. Results

First, one can notice that the initial system performances are very category dependent. At $t = 0$, before any semantic learning, the minimum performances are from 5% to 13% (*cf.* Fig. 2), while the maximum performances are from 52% to 64% (*cf.* Fig. 4). This can be explained by the capabilities of the low-level features to represent well the categories. For instance, *birds* images have very few common colors and textures, while *doors* images have many common features (horizontal and vertical textures). Let us also note that the performance scores we obtain are rather good when the MAP statistic is used.

For all active learners, mean performances increase (*cf.* Fig. 3), but in very different way. This is explained by the behavior of the active learner. As kernel is updated according to the labels at the end of a retrieval process, and these labels depend on the active learner, then the kernel updating also depends on the active learner. For the basic active learner, the curve increases much more after the first sessions than after the last ones. This is certainly because this active method always gives labels in a small part of the space. This allows local semantic clustering, but never gather sparse categories. In the case of the SVM$_{active}$ learner, which chooses unlabeled pictures closest to the SVM margin, the improvement is small. This can be explain by the fact that these active method provide a lot of negative labeled pictures during the first relevance feedback steps, and the semantic learning method is not efficient in such a case. In the case of RETIN AL method, which chooses unlabeled pictures such as training set is balanced, the improvement is significant. Configurations with approximately the same number of positive and negative labels are particularly efficient for the proposed semantic learner.

These results show that the active learning method has a great importance in a semantic learning context. At the end of these experiments ($t = 5000$), the difference between the best and the worst mean performance is 44% (*cf.* Fig. 3). More than improving performances during a single retrieval session, the choice of the active learner has a great influence on the overall results after the semantic learning.

5. CONCLUSION

In this paper, we introduced a semantic learning method RETIN SL to manage the labels provided by users during CBIR experiments. Our approach is based on a kernel matrix in a statistical learning framework. The kernel matrix is updated according to labels provided by users at the end of their retrieval sessions. We introduced two types of receivable algebraic transformations to ensure that the matrix keeps the nice kernel properties. Our technique is designed to model mixed categories and to learn any kind of semantic links.

To handle huge databases, we also proposed a low-rank approximation of the full matrix using a specific factorization and QR decomposition scheme; that enables us to get a complexity linear to the size of the database.

Experiments on a large generalist database show the efficiency of the RETIN SL method with three different active learners. Performances are always improved. These experiments also show the importance of the active learner in the semantic learning context.

We are currently focusing on the building of a complete learning scheme for image retrieval, including local (active learning) and global (semantic learning) aspects.

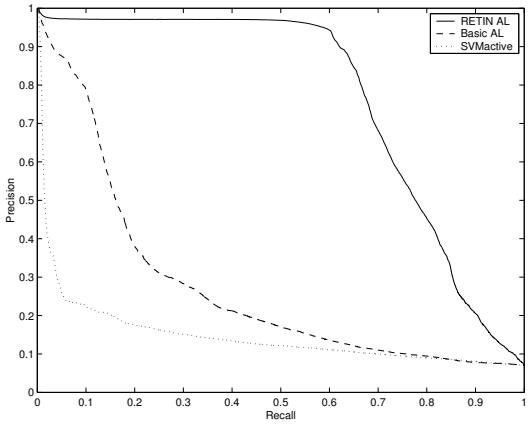

Fig. 1. Precision/Recall curve for the 'savana' category, after 5,000 retrieval sessions.

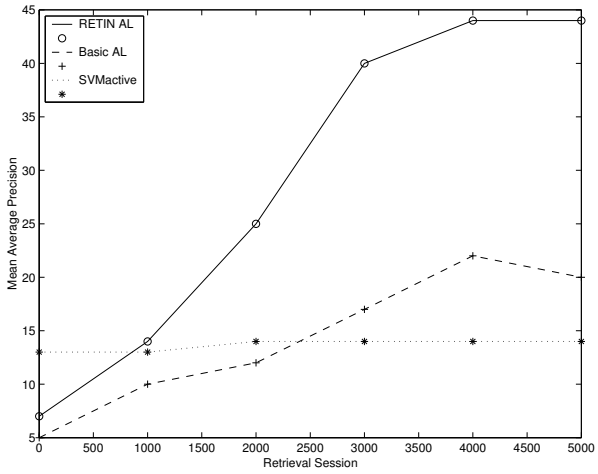

Fig. 2. Min performance for each active learner.

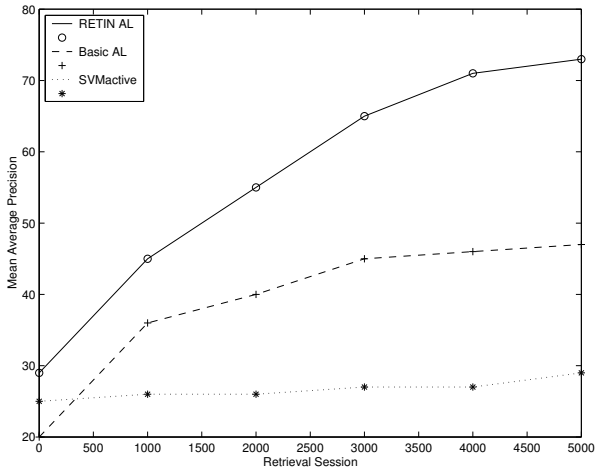

Fig. 3. Mean performance for each active learner.

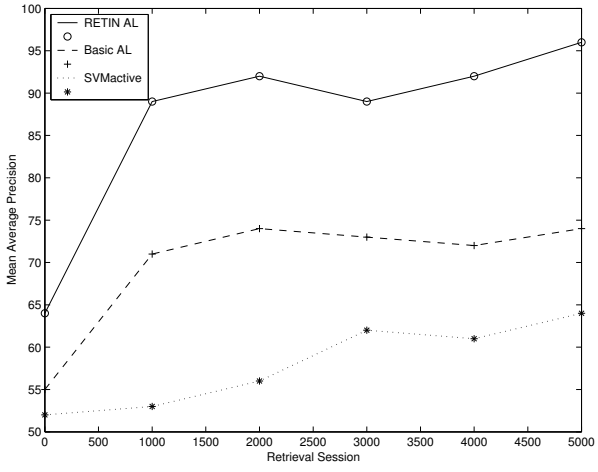

Fig. 4. Max performance for each active learner.

6. REFERENCES

[1] S. Santini, A. Gupta, and R. Jain, "Emergent semantics through interaction in image databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 3, pp. 337–351, 2001.

[2] J.P. Eakins, "Towards intelligent image retrieval," *Pattern Recognition*, vol. 35, pp. 3–14, 2002.

[3] N. Vasconcelos and M. Kunt, "Content-based retrieval from image databases: current solutions and future directions," in *International Conference in Image Processing (ICIP'01)*, Thessaloniki, Greece, October 2001, vol. 3, pp. 6–9.

[4] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *ACM Multimedia*, 2001.

[5] P.H. Gosselin and M. Cord, "RETIN AL: An active learning strategy for image category retrieval," in *IEEE International Conference on Image Processing*, Singapore, October 2004.

[6] D. Cohn, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[7] H. Müller, W. Müller, D. McG. Squire, S. Marchand-Maillet, and T. Pun, "Long-term learning from user behavior in content-based image retrieval," Tech. Rep., Computer Vision Group, University of Geneva, Switzerland, 2000.

[8] D. R. Heisterkamp, "Building a latent semantic index of an image database from patterns of relevance feedback," in *International Conference on Pattern Recognition (ICPR)*, Quebec City, Canada, 2002.

[9] N. Cristianini, J. Show-Taylor, A. Elisseff, and J. Kandola, "On kernel target alignement," in *Neural Information Processing Systems*, Vancouver, Canada, December 2001.

[10] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[11] J. Fournier and M. Cord, "Long-term similarity learning in content-based image retrieval," in *International Conference in Image Processing (ICIP)*, Rochester, New-York, USA, September 2002.