IMAGE RETRIEVAL USING LONG-TERM SEMANTIC LEARNING

Matthieu Cord and Philippe H. Gosselin

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

ABSTRACT

The automatic computation of features for content-based image retrieval still has difficulties to represent the concepts the user has in mind. Whenever an additional learning strategy (such as relevance feedback) can improve the results of the search, the system performances still depend on the representation of the image collection. We introduce in this paper a supervised optimization of a set of feature vectors. According to an incomplete set of partial labels, the method improves the representation of the image collection, even if the size, the number, and the structure of the concepts are unknown. Experiments have been carried out on a large general database in order to validate our approach.

Index Terms— Image classification,Image databases,Learning systems,Information retrieval

1. INTRODUCTION

To manage large image collections, powerful system assistants are required to group images into *clusters* or semantic *concepts*. Most of the time, the low-level features (like color or texture) do not very well match the semantic concepts, and some learning step is usually applied to fill the gap.

If training data are available for each concept, the problem may be solved very efficiently using combinations of classifiers, each of them trained to identify one concept [1]. If not, other approaches use knowledge from user interaction in order to refine the on-line building of concepts. Relevance feedback and active learning [2] increase the system performances, but only during the current retrieval session. Once the session is over, labels are discarded.

The purpose of this paper is to propose learning framework to use all the labels accumulated during previous interactive uses of any retrieval system to improve the feature representation of the images. With such an optimized representation, we attempt to get a better match with semantic concepts. The labels are sampled from a hidden concept that the user has in mind during his retrieval session. Thus, if a large number of labels are available through several retrieval sessions, their combinations should make the hidden concepts stand out.

In order to learn semantic features, some researchers perform a competition of the feature dimensions [3]. Others propose to learn a distance metric [4, 5]. When concepts are very badly represented by features, one can directly focus on the similarities between documents. For instance, in [6], a semantic similarity matrix is computed and stored. The method is relevant to compute semantic links, but has large memory needs. In [7], a clustering of the database is performed to reduce the memory needs and to enhance the system performances. However, the resulting similarities are difficult to exploit with any learning method (classification, active learning, browsing, ...). These strategies usually need specific learning methods, which disable the use of the most powerful ones.

Learning with kernel methods has also been proposed to deal with semantic labels [8, 9]. We recently proposed a kernel matrix updating method, to exploit semantic labels for general database management [10]. However, expressing interesting and efficient data updating rules is not easy when only algebraic transformations on kernels are considered.

To overcome these difficulties, we propose in this paper a new approach working in the feature space, based on a moving of the feature vectors. Our method arranges feature vectors around a set of equidistant concept centers, without an explicit computation of those centers. For the equidistance property, we introduce a theorem that let us compute all the feature movements.

According to an incomplete set of partial labels, the method improves the representation of the image collection, even if the size, the number and the structure of the concepts are unknown. Contrary to [8, 11], the method may learn many concepts with mixed information. Moreover, in opposition to $O(N^2)$ methods like [5], the complexity of our technique is no more dependent on the database size, but only on the label set size.

2. WEAKLY SUPERVISED LEARNING FRAMEWORK

The problem addressed in this paper is a particular learning problem, because of the nature of the training set. Let us note $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ the whole set of documents represented by $\mathbf{x}_i \in \mathbb{R}^p$. Suppose that the documents are gathered in a finite (but unknown) number of concepts, and these concepts do not necessarily form a clustering. Thus, an image can belong to several concepts. For instance, one can find buildings, cars, houses, or landscape, but also cars in front of a building or a house, or houses in a landscape.

A usual training set in classification problems is a set of couples (data,concept). In this paper, we consider a different learning framework. Indeed, let us consider that the training set is composed of a set $\mathbf{Y}$ of vectors $\mathbf{y}_p \in [-1, 1]^n$: each $\mathbf{y}_p$ is a partial labeling of the set $\mathbf{X}$, according to one of the hidden concepts[1]. This type of training data is very common in information retrieval framework.

Learning from such a training set $\mathbf{Y}$ is far from being trivial, as the knownledge of the concept associated with each training sample is unknown. For instance, a training sample can be "$\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$ are in *some* concept, and $\mathbf{x}_l$ is not in that concept". Furthermore, we do not assume that a training sample is large enough to build the whole associated concept, which makes the learning problem $weakly$-supervised.

[1]Every positive (resp. negative) value y_{ip} means that the image represented by $\mathbf{x}_i$ is (resp. not) in this concept, as much as y_{ip} is close to 1 (resp. -1). Every value y_{ip} close to zero means that there is no information for $\mathbf{x}_i$ about this concept.

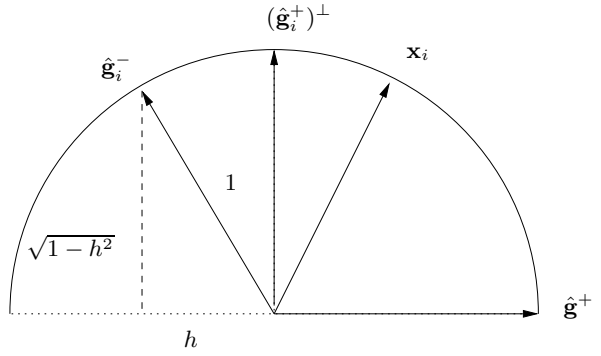

Fig. 1. Estimated center $\hat{\mathbf{g}}_i^-$ for a negative labeled vector $\mathbf{x}_i$, relatively to the positive concept center $\hat{\mathbf{g}}^+$.

The challenge is to use this set of partial labeling in order to learn the concepts.

3. CONCEPT VECTOR LEARNING METHOD

We propose a vector-based approach which arranges vectors in $\mathbf{X}$ around a set of concept centers $\mathbf{g}_j$, without explicitly compute the $\mathbf{g}_j$. The idea is to build a new set $\mathbf{X}^\star$ of vectors $\mathbf{x}_i^\star$ such as the vectors are clustered by concepts.

The main difficulty is to build these clusters in the weakly supervised framework previously described. We propose an adaptive scheme using $\mathbf{y}_p$ one after another, shifting the corresponding labeled vector $\mathbf{x}_i$ ($y_{ip} \neq 0$). The idea is to move positive labeled vectors towards an estimation $\hat{\mathbf{g}}$ of the concept center $\mathbf{g}$ of the cluster those vectors may be in, and to move away from $\hat{\mathbf{g}}$ the negative labeled vectors. The problem is the estimation $\hat{\mathbf{g}}$ of a concept center $\mathbf{g}$, according to a $\mathbf{y}_p$. For positive labeled vectors, we propose to move them towards a single estimated center corresponding to the barycenter. To shift the negative labeled vectors, we propose a different strategy by considering several potential centers. We propose a theorem that offers us an effective solution for the corresponding move of the negative vectors. This is the most original part of this work that we justify and comment in the following sections.

3.1. Global scheme

Vectors $\mathbf{y}$ are randomly sampled from the whole set $\mathbf{Y}$, and $\mathbf{X}$ is updated: we compute an estimated concept center $\hat{\mathbf{g}}_i$ for each of the labeled vectors $\mathbf{x}_i$ ($y_i \neq 0$) . Next, we move the labeled vectors towards their corresponding centers:

$$\forall i \in 1..n,\ \mathbf{x}_i \leftarrow \mathbf{x}_i + \rho|y_i|(\hat{\mathbf{g}}_i - \mathbf{x}_i)$$

Repeating this update several times with decreasing ρ, the set $\mathbf{X}$ converges to a set $\mathbf{X}^\star$. In the case of an efficient algorithm, vectors in $\mathbf{X}^\star$ are in clusters around the true concept centers $\mathbf{g}_j$.

3.2. Center computing

Positive labels in $\mathbf{y}$ mean that the corresponding vectors are in the same concept, so we propose to compute the barycenter of the labeled vectors:

$$\hat{\mathbf{g}}^+ = \frac{1}{\sum_j |y_j|} \sum_j y_j \mathbf{x}_j$$

Then each positive labeled vector $\mathbf{x}_i$ moves towards the vector $\hat{\mathbf{g}}_i = \hat{\mathbf{g}}^+$.

Negative labels in $\mathbf{y}$ mean that the corresponding vectors are not in the concept. This does not mean that all negative labeled vectors are in the same concept. It means that the negative labeled vectors are not around the possible center $\hat{\mathbf{g}}^+$ of positive labeled vectors.

To be able to propose an effective tuning for negative vectors, we introduce the following constraint on the concept center distribution: we force them to be equidistant. This property makes sense as soon as we do not have any *prior* about the distribution of these semantic concepts in the feature space. Additionally, it offers a very interesting property to set or move vectors between two centers without changing their distances to other centers. To use this constraint in the estimation of the concept center $\hat{\mathbf{g}}_i^-$ for each negative labeled vector $\mathbf{x}_i$, we have established the following theorem:

Let $\mathbf{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_q\}$ be a set of different vectors $\mathbf{g}_j \in \mathbb{R}^{q-1}$ such as $\forall j = 1..q,\ ||\mathbf{g}_j|| = 1$. Then the vectors of $\mathbf{G}$ are equidistants if and only if their mutual distance is $d = \sqrt{2(1 + \frac{1}{q-1})}$. (See appendix for proof).

So, the only way to get equidistant centers (for q centers of dimension $q-1$) is to fix the distance from one to another to d. This property gives us an effective solution to compute the negative vector update, by setting the negative centers $\hat{\mathbf{g}}_i^-$ to the distance d from the positive center $\hat{\mathbf{g}}^+$. In this scope, we choose $\hat{\mathbf{g}}^-$ in the plan spanned by $\mathbf{x}_i$ and $\hat{\mathbf{g}}^+$ such as its distance to $\hat{\mathbf{g}}^+$ is d (*cf.* Fig. 1).

A basis of this plan is $\left(\hat{\mathbf{g}}^+, (\hat{\mathbf{g}}_i^+)^\perp\right)$ where

$$(\hat{\mathbf{g}}_i^+)^\perp = \frac{\mathbf{x}_i - \langle \mathbf{x}_i, \hat{\mathbf{g}}^+ \rangle \hat{\mathbf{g}}^+}{||\mathbf{x}_i - \langle \mathbf{x}_i, \hat{\mathbf{g}}^+ \rangle \hat{\mathbf{g}}^+||}$$

Next, as we need to have $||\hat{\mathbf{g}}^+ - \hat{\mathbf{g}}_i^-|| = d$, then $\langle \hat{\mathbf{g}}^+, \hat{\mathbf{g}}_i^- \rangle = h = -\frac{1}{q-1}$, and:

$$\hat{\mathbf{g}}_i^- = h\hat{\mathbf{g}}^+ + \sqrt{1-h^2}(\hat{\mathbf{g}}_i^+)^\perp$$

By this way, all negative labeled vector $\mathbf{x}_i$ move towards the vector $\hat{\mathbf{g}}_i = \hat{\mathbf{g}}_i^-$.

4. EXPERIMENTS

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 50 concepts. Each concept is built from 2 or 3 of the COREL folders. The concept sizes are from 50 to 300. The set of all the concepts covers the whole database, and many of them share common images.

We randomly built a set of labels Y simulating the use of the system. For instance, this set could be made from the labels given by users during the use of an image retrieval system. This set could also be made from text associated with each image. In all cases, we assume that the labels are incomplete, and have few non-zero values. In this context, we build partial labeling $\mathbf{y}_p$ with 50 positive values and 50 negative values. We next train the concept vector learning algorithm from 100 to 500 $\mathbf{y}_p$ training samples.

In order to evaluate the improvement, we experimented an optimized SVM classification[12], with a training set of 100 labels. As the size of concepts (100 to 300) is small against the size of the

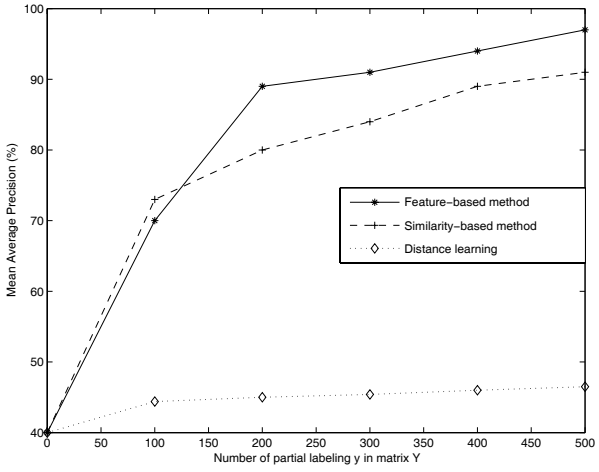

Fig. 2. Mean Average Precision error according to the number of partial labeling.

database (6,000), we used the Mean Average Precision[2] to evaluate the performances. In this case, the error of classification is, in these cases, less relevant for comparison. Figure 2 shows the results for three different methods : a distance learning method [5], a similarity-based learning method [10] and the feature-based method proposed in this paper. The performances quickly increase with few partial labeling, and stabilize themselves with more labeling. The distance learning method does not improve a lot the performances, certainly because the concept are mixed. The feature-based method improves the most the performances, and furthermore is faster than the similarity-based method. A few second are required for optimization with the method proposed in this paper, whereas the similarity-based method, which works on an eigendecomposition of the Gram matrix, requires several minutes.

We show in Fig. 3–6 an example of retrieval before and after semantic learning. In both cases, the user is looking for mountains, and the query is composed of two positive examples (the images with a small green square in figures). Before optimization, there is already irrelevant pictures within the closest pictures to the query (*cf.* Fig. 3). After optimization, since users have labeled mountains as being in the same concept during the previous sessions, the closest images are all mountains (*cf.* Fig. 4–6).

5. CONCLUSION

In this paper, we introduced a concept vector learning method which improves the representation of a document collection, with few constraints on training data. The method is mostly based on the equidistance of concept centers, gathering the vectors of the same concept around each corresponding centers, and distributing the vectors in several concepts between these centers. Thus, the method is able to deal with mixed concepts, with the only constraint that the dimension of the vectors must be larger than the number of concepts. We are actually working on an extension of this method to overcome this constraint, by applying the optimization in an infinite space, with a framework similar to the quasiconformal kernels approach [13, 14]. If we assume that images need to be gathered into concepts, then the method deals with a context where the size, the number and the structure of the concepts are unknown. Experiments carried out on real data demonstrate the efficiency of the method.

[2]cf. TREC VIDEO conference: http://www-nlpir.nist.gov/projects/trecvid/

A. PROOF OF THEROEM 1

Assume that we have a set $\mathbf{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_q\}$ of normalized and different vectors of dimension $q-1$, with the same distance d from one to another.

Then $\forall j, j' \in 1..q,\ \langle \mathbf{g}_j, \mathbf{g}_{j'} \rangle = h = 1 - \frac{d^2}{2}$.

Let $\mathbf{K} = \mathbf{G}^\top \mathbf{G}$ be the $q \times q$ matrix of all dot products between all vectors of $\mathbf{G}$.

Then $\mathbf{K}$ is 1 on the diagonal, h otherwise. $\mathbf{K}$ is the dot product matrix of q vectors of dimension $q-1$, then $\det \mathbf{K} = 0$. In order to compute the determinant of $\mathbf{K}$, we compute the characteristic polynomial of $\mathbf{K}$: $\det(\mathbf{K} - \lambda \mathbf{Id})$.

The matrix $\mathbf{K} - \lambda \mathbf{Id}$ is $(1-\lambda)$ on the diagonal, h otherwise. If we set $\lambda = \lambda' - h + 1$, then $\mathbf{K} - \lambda \mathbf{Id} = h\mathbf{e}\mathbf{e}^\top - \lambda' \mathbf{Id}$, with $\mathbf{e}^\top = (1 \ldots 1)$.

The characteristic polynomial of the rank one matrix $h\mathbf{e}\mathbf{e}^\top$ is $(-\lambda')^{q-1}(hq - \lambda')$.

As $\lambda' = h - 1 + \lambda$, then $\det(\mathbf{K} - \lambda \mathbf{Id}) = ((1-h)-\lambda)^{q-1}((1+(q-1)h)-\lambda)$ so that: $\det(\mathbf{K}) = ((1-h))^{q-1}((1+(q-1)h)) = 0$.

As $\mathbf{g}_j \neq \mathbf{g}_{j'}$, then $h \neq 1$. It follows that $(1+(q-1)h) = 0$, *i.e.* $d^2 = 2(1 + \frac{1}{q-1})$.

B. REFERENCES

[1] A. Natsev, M. Naphade, C.Y. Lin, and J.R. Smith, "Overcomplete representation and fusion for semantic concept detection," in *IEEE International Conference on Image Processing (ICIP)*, Singapore, October 2004.

[2] E. Chang, B. T. Li, G. Wu, and K.S. Goh, "Statistical learning for effective visual information retrieval," in *IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

[3] 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.

[4] M Schultz and T Joachims, "Learning a distance metric from relative comparisons," in *Neural Information Processing Systems*, 2003.

[5] E P Xing, A Y Ng, M I Jordan, and S Russell, "Distance metric learning, with application to clustering with side-information," in *Neural Information Processing Systems*, Vancouver, Bristish Columbia, December 2002.

[6] 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.

[7] Junwei Han, Mingjing Li, Hongjiang Zhang, and Lei Guo, "A memorization learning model for image retrieval," in *IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

[8] Nello Cristianini, John Shawe-Taylor, A Elisseff, and J Kandola, "On kernel target alignement," in *Neural Information Processing Systems*, Vancouver, Canada, December 2001.

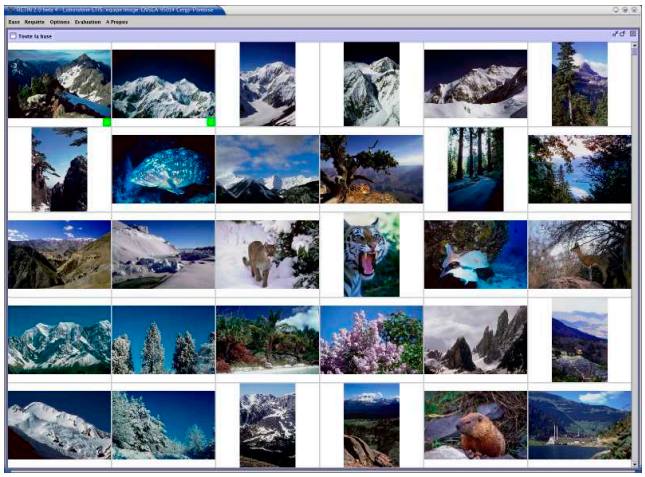

Fig. 3. 30 closest pictures to the query, before semantic learning.

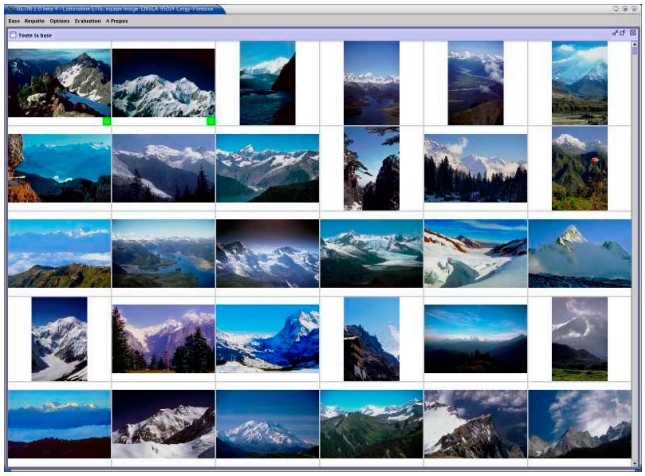

Fig. 4. 30 closest pictures to the query, after semantic learning.

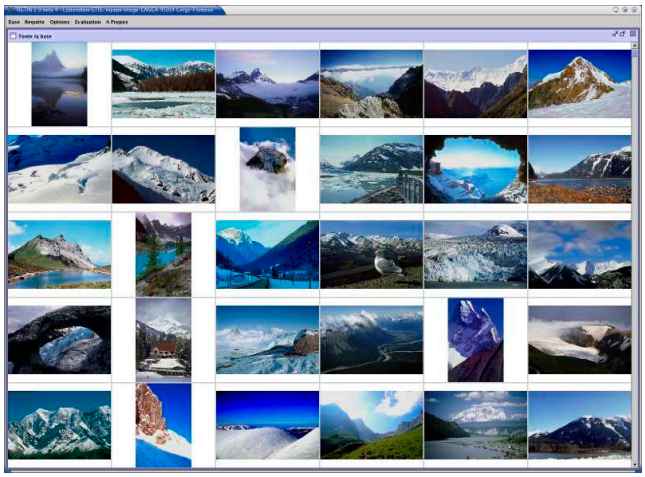

Fig. 5. 31-60 closest pictures to the query, after semantic learning.

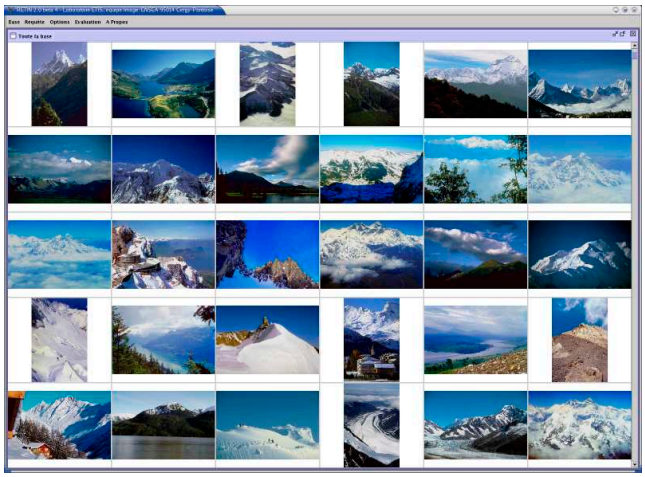

Fig. 6. 61-90 closest pictures to the query, after semantic learning.

[9] Douglas R Heisterkamp, "Building a latent semantic index of an image database from patterns of relevance feedback," in *International Conference on Pattern Recognition*, Quebec City, Canada, 2002, pp. (4):132–137.

[10] Philippe Henri Gosselin and Matthieu Cord, "Semantic kernel learning for interactive image retrieval," in *IEEE International Conference on Image Processing*, Genova, Italy, September 2005.

[11] G R G Lanckriet, Nello Cristianini, N Bartlett, L El Ghaoui, and M I Jordan, "Learning the kernel matrix with semi-definite programming," in *International Conference on Machine Learning*, Sydney, Australia, 2002.

[12] Philippe Henri Gosselin and Matthieu Cord, "RETIN AL: An active learning strategy for image category retrieval," in *IEEE International Conference on Image Processing*, Singapore, October 2004.

[13] S Amari and S Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, pp. 12(6):783–789, 1999.

[14] Douglas R Heisterkamp, Jing Peng, and H K Dai, "Adaptive quasiconformal kernel metric for image retrieval," in *International Conference on Computer Vision and Pattern Recognition*, 2001.