A kernel-based active learning strategy for content-based image retrieval

I. Daoudi, K. Idrissi
Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
{dimane, kidrissi}@liris.cnrs.fr

Abstract

Active learning methods have attracted many researchers in the content-based image retrieval (CBIR) community. In this paper, we propose an efficient kernel-based active learning strategy to improve the retrieval performance of CBIR systems using class probability distributions. The proposed method learns for each class a nonlinear kernel which transforms the original feature space into a more effective one. The distances between user’s request and database images are then learned and computed in the kernel space. Experimental results show that the proposed kernel-based active learning approach not only improves the retrieval performances of kernel distance without learning, but also outperforms other kernel metric learning methods.

1. Introduction

Content-based image retrieval has received much interest in the last decades due to the large digital storage and easy access to images on computers and through the World Wide Web [1]. A common scheme used in CBIR is to first automatically extract from images a set of features (color, texture, shape, etc.) structured into descriptors (indexes). These indexes are then used in a search engine to compare, classify, rank, etc. the database content.

The two determining factors for image retrieval performances are, on one hand, the considered features to describe the images, and on the other, the distance used to measure the similarity between a query and images in the database. It is well known that for a specific set of features, the performance of a content-based image retrieval system depends critically on the similarity or dissimilarity measure in use. Learning can be considered as one of the most interesting way to reduce the semantic gap and also to improve the performances of CBIR systems.

Different learning strategies, such as supervised learning, unsupervised learning, semi-supervised learning and active learning may be considered to boost the efficiency of information retrieval system.

In the supervised distance metric learning, all training data are labeled and cast into pairwise constraints: the equivalence constraints where pairs of data belong to the same classes, and the inequivalence constraints where pairs of data belong to different classes. The supervised distance metric learning can be divided into two categories: In the first one, the distance metric is learned in a global sense, i.e., to satisfy all the pairwise constraints simultaneously. A review of various learning methods of this category can be found in [2]. In the second category, distance metric is learned in a local sense, satisfying only local pairwise constraints. Several authors [3], [4], used this approach to learn appropriate distance metrics for \( k-NN \) classifier. Particularly, in [5], a Quasiconformal Kernel for nearest neighbor classification is proposed which adjusts the Radial Basis function by introducing weights based on both local consistency of class labels and labeling uncertainty. In [6], the authors propose a technique that computes a locally flexible distance metric using SVM. As proposed in [7] and [5], Bermejo et al [6] attribute some weights to the features, based on their relevance to the class conditional probabilities for each query. In [8], Hoi et al. propose a simple and efficient algorithm to learn a full ranked Mahalanobis distance metric. This approach constructs a metric in kernel space, based on a weighted sum of class covariance matrices.

The unsupervised distance metric learning used the unlabeled data for training. The main idea of this approach is to learn low-dimensional manifold where distances between most of observed data are preserved. It can be divided into nonlinear and linear approaches. The most popular methods for nonlinear unsupervised dimensionality reduction are ISOMAP [9], Locally Linear Embeding (LLE) [10], and Laplacian Eigenmap (LE) [11]. ISOMAP preserves the geodesic distances between any two data points, while LLE and LE focus on the preservation of the local neighbor structure. The well-known algorithms for the unsupervised linear methods are the Principal Component Analysis (PCA)
Semi-supervised learning refers to the use of both labeled and unlabeled data for training. For this of methods, emerging distance metric learning techniques are proposed. For example, Relevance Component Analysis (RCA) learns a global linear transformation by using only the equivalent constraint [15]. Discriminate Component Analysis (DCA) improves the RCA by incorporating the negative constraints [8]. More recently, Hong et al. proposed a kernel-based distance metric learning for content-based image retrieval [16]. Other methods [17], propose a joint learning of labels and distance metric approach, which is able to simultaneously address the two difficulties.

Active learning methods are an extension of the semi-supervised distance metric learning, they are considered as one of the most interesting techniques for improving the CBIR performance. The challenge of those methods is how to identify the most informative unlabeled data such that the retrieval performance could be improved most efficiently. In this paper, we introduce a novel scheme of distance metric active learning for content-based image retrieval. A good distance metric would lead to tight and well-separated clusters in the projected space. In our idea, this can be quantified by the use of a new criterion, which is the ratio between class probabilities of the vectors that are respectively different and similar to the query. The criterion resembles the one used in Adaptive Quasiconformal Kernel (AQK) [5], except that we compute a metric learning using an active learning setting, while AQK assumes that labels are already known. The proposed method maps data vectors into a kernel space and learns with user’s interaction, relevant and irrelevant features’ vectors using class probability distributions. Based on quasiconformal transformed kernels, the proposed active learning process generates for each class a suitable similarity model by accumulating classification knowledge collected over multiple query sessions. Unlike the most of existing learning metric distance, the proposed method is well adapted to the content-based image retrieval where data are generally composed by heterogeneous features (color, shape, texture). This is due to these multifold advantages. 1) The problem of the adaptability of the used metric distance (kernel function) to the image database can be tackled. 2) A good distance metric for each class can be constructed with only few training samples. 3) No weighting parameters are needed since the algorithm automatically determines the metric distance. Experimental tests show encouraging results in comparison of the state-of-the-art.

The next section presents the kernel-based similarity model used in this paper. The proposed active distance metric learning strategy is described in section 3. Section 4 deals with our experimental results before concluding.

2. Kernel-based similarity model

We propose in this paper a nonlinear similarity measure based on kernel function. The approach consists in finding a mapping \( \Phi \) from an input space \( I \) to a Hilbert space \( \chi \). Such a mapping should verify \( \langle \Phi(a), \Phi(b) \rangle = k(a,b) \), so the inner product between two vectors \( a \) and \( b \), can be considered as a measure of their similarity. Therefore the distance between \( a \) and \( b \) is defined as:

\[
dist(a,b) = \langle \Phi(a), \Phi(b) \rangle = k(a,a) - 2k(a,b) + k(b,b) \quad (1)
\]

The idea of kernel function framework is to deal with the dot product \( \langle \Phi(a), \Phi(b) \rangle \) which is evaluated by \( k(a,b) \) in the original space, instead of the mapped vectors \( \Phi(a) \) and \( \Phi(b) \). This allows us to define efficient learning scheme in the Hilbert space. In our CBIR context, we are working on heterogeneous features, thus, an interesting kernel function is the Gaussian (GRBF) one:

\[
k(a,b) = \exp\left(-\frac{d^2}{2\delta^2}\right) \quad (2)
\]

where \( \delta \) is a scaling parameter and \( a \) and \( b \) are two vectors in the input space. We propose in the next section an efficient strategy to compute the optimal value of the scaling parameter for our application.

Another interest of the kernel framework in distance metric learning context is the ability to create a new kernel function derived from the existing ones depending on the considered application [5]. We exploit this property to modify the similarity measure, by modifying the kernel, and using an active learning scheme. The new kernel function is defined as:

\[
\tilde{k}(a,b) = c(a)c(b)k(a,b) \quad (3)
\]

where \( c(a) \) is a positive function (detailed in 3.3).

3. Active distance metric learning scheme

The proposed active distance metric learning comprises three steps. We first map the input vectors into a feature space using Kernel Principal Component Analysis (KPCA) [9], and, in the second step, the best parameters of the KPCA are estimated to well-separated clusters in the new space and finally, the similarity model is learned from data.
3.1 Step 1: Kernel Principal Component Analysis (KPCA)

Let \( x_i (i = 1, ... , N) \) be \( N \) vectors in the input space \( I \), and \( \Phi(x_i) \) their nonlinear mapping into a feature space \( \chi \). KPCA finds the principal axes by diagonalizing the covariance matrix:

\[
C = \frac{1}{N} \sum_{i=1}^{N} \Phi(x_i)\Phi(x_i)^T
\]

(4)

where \( \Phi(x_i)^T \) is the transpose of \( \Phi(x_i) \). The principal orthogonal axes \( V_i (i = 1, ... , M) \) (\( M \) is the dimensionality of the feature space) can be found by solving the eigenproblem:

\[
\lambda_i V_i = C V_i
\]

(5)

where \( V_i \) and \( \lambda_i \) are respectively the \( i \)th eigenvector and its corresponding eigenvalue. It can be shown [9] that the solution of the above eigenproblem lies in the span of the data, i.e:

\[
\forall p = 1, ... , M, \exists \alpha_{p1}, ..., \alpha_{pN} \in \mathbb{R} \ s.t. \ V_p = \sum_{j=1}^{N} \alpha_{pj} \Phi(x_j)
\]

(6)

where \( \alpha_p = (\alpha_{p1}, ..., \alpha_{pN}) \) are found by solving the eigenproblem [9]:

\[
K\alpha_p = \lambda_p \alpha_p
\]

(7)

where \( K \) is the Gram matrix defined by:

\[
k_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle \quad \text{where} \quad i, j = 1, ..., N \quad \text{are respectively the row and the column indices of} \ K \ .\]

The eigenvectors \( V_i \) are sorted in a decreasing order according to the magnitudes of their corresponding eigenvalues \( \lambda_i \), and the first eigenvectors are selected as the basis and are used to map data vectors into the feature space. For any input vector \( x_p \), the \( i \)th principal component \( \tilde{y}_i \) of \( \Phi(x_p) \) is given by:

\[
\tilde{y}_i = \frac{1}{\sqrt{k_i}} \sum_{j=1}^{N} \alpha_{j} < \Phi(x_j), \Phi(x_p)> \tag{8}
\]

In KPCA, the nonlinear mapping \( \Phi(x) \) into the kernel space is determined by the nonlinear function \( k \) (GRBF in our case). In the next section, we show that for different given values of \( \delta \) and \( d \), where \( \delta \) is the scaling parameter of the GRBF and \( d \) is the dimensionality of the feature space, we obtain a set of functions which provide various approximation of the original data nonlinearity in the feature space. The optimal kernel function obtained is then used to build a similarity measure.

3.2 Step 2: KPCA parameters estimation

As described previously, KPCA deals with nonlinear transformation via nonlinear kernel functions. In the used kernel function (GRBF), there are two parameters \( d \) and \( \delta \) that must be predetermined, knowing that they have significant impact on image representation in feature space. Ideally, compact and informative image representation will facilitate the retrieval process. An illustrative example is given in the figure 1, to show the influence of these two parameters on data structure and the classification task. For this, a set of three classes is used; each one consists of 40 vectors with dimensionality 15. The first image on each row represents the original data (1.a and 1.e). In the first row, we present the first and the second principal components obtained by KPCA using for \( \delta \), from left to right respectively, the values 0.01, 2 and 20, and using a fixed value of \( d = 2 \). We can see from these figures that the kernel parameter \( \delta \) has a significant effect on the class separability. When increasing \( \delta \) until a certain value, better separation of the class vectors is obtained (1.e). In our case, for \( \delta = 2 \) the best separation is reached, while for a large value of \( \delta \) (in our case \( \delta = 20 \)), 2 classes are obtained instead of 3 classes. In the second row, we fix \( \delta = 1 \) and we plot the three eigenvectors, obtained from KPCA, corresponding to the three largest eigenvalues. Thus, we have the 2nd, 3rd and 4th eigenvectors in figure 1.f (with \( d = 4 \)), the 5th, 6th

Figure 1. Principal Components for different values of \( \delta \) and \( d \)
and 7th eigenvectors in figure 1.g (with \( d = 7 \)), and
the 8th, 9th, and 10th eigenvectors in figure 1.h (with \( d = 10 \)). We see that the top principal eigenvectors (figure 1.f and 1.g), capture the major data variance allowing a better data representation, while the remaining ones (figure 1.h) correspond to the less significant variance. Finally the choice of \( d \) and \( \delta \) values is crucial as it can widely influence the success or the failure of the retrieval.

Our idea is to find a good distance metric which can be used not only to measure similarity between data, but also to propose a new representation of them. We propose in this investigation a learning strategy for the parameters \( d \) and \( \delta \), which allow obtaining maximum of both class separability and statistical variance. The goal of the first condition is to find a nonlinear transformation that leads to an optimal distance metric which allows to maximize the inter-class variance and to minimize the intra-class variance in the feature space. Therefore, it offers an effective data representation that supports more accurate search strategies. This condition can be evaluated by the class separability criterion defined as follows:

Let \( \Phi(x_i) (i = 1,...,N) \) be the \( N \) vectors in the feature space, \( l_i \) be the number of vectors (descriptors) that belong to class \( i \), \( m_i = \frac{1}{l_i} \sum_{j=1}^{l_i} \Phi(x_j) \) be the mean vector of class \( i \) in the feature space, and
\[
m = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{l_i} \Phi(x_j),
\]
the mean vector of all vectors \( N \).

The average within-cluster distances \( S_w \), which correspond to intra-class variance, and the average between-clusters distances \( S_b \), which correspond to the inter-class variance, can be calculated by:
\[
S_w = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{l_i} (m_i - m)(m_i - m)^T,
\]
\[
S_b = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{l_i} (\Phi(x_i) - m)(\Phi(x_i) - m)^T.
\]

The idea is to find the non linear transformation (kernel function which depends on \( d \) and \( \delta \)) that maximizes the following criterion:
\[
\gamma = S_b / S_w \quad (9)
\]

The goal of the second condition is to select the value of the dimensionality \( d \), corresponding to the number of eigenvectors which capture the major statistical variation in the data. In our case, we select \( d \) eigenvectors in order to capture 98% of the variance, i.e:
\[
\rho = \sum_{k=4}^{d} \sum_{k=4}^{d} \geq 0.98 \quad (10)
\]

To find the values of \( d \) and \( \delta \) that satisfy the two criteria defined previously, we propose to project the original data in a feature space with KPCA using different values of \( d \) and \( \delta \). An empirical selection is used to determine the initial values of \( d \) and \( \delta \).

In a first step, we select all the couples whose \( \gamma \) value verifies:
\[
\gamma_{d,\delta} \geq 98\% \quad (11)
\]
In the second step, we select, among all the couples selected in the first step, those whose statistical variance \( \rho \) verifies:
\[
\rho_{d,\delta} \geq 98\% \quad (12)
\]
Finally we keep among the obtained couples those having the lowest value of \( d \) and \( \delta \). Experimental results (\& 4.1) show that the learning strategy used to select the values of \((d,\delta)\) achieved the highest retrieval performances.

### 3.3 Step 3: Active learning mechanism

In this step, the non-linear similarity models are learned from user feedback and \( k - NN \) search is then applied based on those models. Note that the first step of the retrieval process (ie before applying the metric learning process) is based on the GRBF, using the optimal kernel parameters that we have been found through the strategy described in subsection 3.2.

To perform our active distance metric learning strategy, we create a new kernel \( k \) form the previous one (Equation (3)), and hence a new kernel distance:
\[
dist(a,b)^2 = k(a,a) - 2k(a,b) + k(b,b)
\]
\[
= (a)^2 k(a,a) - 2(a)(b)k(a,b) + (b)^2 k(b,b)
\]
The goal is to create for each class, a kernel function that expand the distance around descriptors whose class probability distributions are different from the query and contract the distance around descriptors whose class probability distribution is similar to the query. This can be achieved by the use of adaptive quasi-conformal kernel (AKQ) \[5\] which create a new feature space, where relevant vectors are brought closer to each other, and irrelevant ones are moved far from relevant vectors. Based on this approach \( c(x) \) can be computed as follows:
\[
c(x) = \frac{P(x/D)}{P(x/S)} \quad (14)
\]
where \( P(x/D) = \frac{1}{\sum_{d} S_d} \sum_{d} \frac{1}{S_d} \) and \( P(x/S) = \frac{1}{\sum_{S} d} \sum_{S} \frac{1}{d} \)

\([S]\) denotes the number of similar descriptors, and \([D]\) denotes the number of dissimilar descriptors. The set of similar \( S \) and dissimilar \( D \) images are used to create a suitable similarity model. They can be computed as follows: a set of \( w \) images are randomly selected from each class so as to be used as queries. As described in figure (2), for each query \( Q_i \) (\( i = 1 \) to \( w \)),
a two step mechanism is processed. First, the system returns the top k − NN images using the similarity model defined by Equation (2), and based on the user interaction, the system identifies \( S_i = \{s_{i,1}...s_{i,N}\} \) and \( D_i = \{d_{i,1}...d_{i,N}\} \) respectively as the set of similar and dissimilar images. The second step consists of selecting from the set \( S_i, N_b \) images that will be used as query stimulating the k − NN search and producing two new sets \( \{S_{i,1}...S_{i,NB}\} \) and \( \{D_{i,1}...D_{i,NB}\} \) of similar and dissimilar images. Finally, we define the sets \( S \) and \( D \) as: \( S = \{s_{i,j}\}_{j=1}^{1,NB}, D = \{d_{i,j}\}_{j=1}^{1,NB} \). Then we compute the suitable similarity model according to equations (13) and (14) for an efficient image retrieval.

4. Experimental Results

Experiments are based on the well-known Coil-100 image database of Columbia University [18]. It contains 7200 images belonging to 100 different classes. Each class contains 72 objects generated by rotating the object at an interval of 5 degree. To describe Coil-100 database images, we use color and shape descriptors because they are well adapted to this database. For color descriptor, we use LAB histogram [18] quantized upon 192 bins, RGB dominant colors, spatial coherency, and percentage of colors [20] upon a vector of 25 bins. Angular Radial Transform (ART) [21] is used as shape descriptor, which is well adapted to Coil-100 database, as each image contains one single object. The final image descriptor is a vector of 252 components (217 for color and 35 for shape). Two experiments are conducted, the first one deals with KPCA parameters estimation strategy, and the second one evaluate the active distance metric learning.

4.1 Kernel distance evaluation

To evaluate the performance of the proposed strategy, the recall and precision parameters are used. We first apply the learning strategy described previously to find the best value of \((d, \delta)\). Thus, we build the optimal kernel function (GRBF), which allows not only to best approximate the non-linearity in the feature space using KPCA, but also to best measure the similarity between two vectors and therefore, between their corresponding images. In our tests, the optimal values are \((d=6, \delta=1)\). In the second experiment, we have compared the similarity search quality using the kernel function for different values of the couple \((d, \delta)\) (optimal and non optimal). The retrieval quality was also compared with the use of Euclidian distance. The comparison results in terms of average recall and precision are given in figure 3. We can see that different kernel metric parameters values involve different retrieval performances and the best results are obtained, as expected, when the optimal parameters values are used. We can also notice that for particular values of the kernel parameters \((d=55, \delta=12)\), the performances of Euclidian and Kernel approaches are similar, and therefore the corresponding curves overlap.

4.2 Active learning strategy evaluations

In this experiment, we compare the retrieval performances obtained with our active distance metric learning and those obtained when using kernel distance without learning. Our approach is also compared to Mahalanobis distance learning with kernel DCA [8]. This method is based on the maximization of the intra-class variance and the minimization of inter-class variance to learn Mahalanobis distance which is classically used as reference in many papers to compare distance learning performances. To measure the retrieval performance, a set of 600 images are randomly selected from the database and are used as queries. Figure 4(a) shows the retrieval results on the coil-100 database. We can see that our active distance
metric learning improves significantly the retrieval performance and outperforms kernel metric and Mahalanobis metric learning with DCA.

Another scenario to compare the image retrieval performance is to use metric learning. We split Coil-100 database into two sets, 80% of total images for training and 20% for testing. Figure 4 (b) presents the retrieval results, and we can see that our method still outperforms kernel metric, and Mahalanobis metric learning with KDCA.

5. Conclusion

In this paper, we have proposed an efficient active distance metric learning method to boost the retrieval performance continuously by accumulating classification knowledge collected over multiple query session. Not only does our method based on quasiconformal kernels improve the retrieval performance of kernel distance, it also outperforms the Mahalanobis metric learning with kernel DCA due to its higher flexibility in metric learning.

6. References

[18] www1.cs.columbia.edu/CAVE/software/softlib/coil-100