A SEMI-SUPERVISED METRIC LEARNING FOR CONTENT-BASED IMAGE RETRIEVAL

I. Daoudi[1,2], K. Idrissi[1], S. Ouatik[3]

[1]Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
[2]Faculté Des Sciences, UFR IT, Université Mohamed V-Agdal BP. 1014, Rabat, Maroc
[3]Laboratoire d'Informatique, Statistiques et Qualité, LISQ, Faculté Des Sciences, Fès, Maroc
{dimane, kidrissi}@liris.cnrs.fr, s_ouatik@yahoo.com

ABSTRACT

In this paper, we propose a kernel-based approach to improve the retrieval performance of CBIR systems by learning a distance metric based on class probability distributions. Unlike other metric learning methods which are based on local or global constraints, the proposed method learns for each class a nonlinear kernel which transforms the original feature space to a more effective one. The distances between query and database images are then measured in the new space. Experimental results show that our kernel-based approach not only improves the retrieval performances of kernel distance without learning, but also outperforms other kernel metric learning methods.

Index Terms— Similarity search, kernel functions, CBIR, k nearest neighbor search

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 hand 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. Distance learning can be considered as one of the most interesting issue to improve the performances of CBIR systems and also to reduce the semantic gap.

Different learning strategies, such as supervised, unsupervised and semi-supervised distance metric learning are used to define a suitable similarity measurement for content-based image retrieval.

The supervised approach 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 approach, 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 then some weights to the features, based on their relevance to the class conditional probabilities for each query. Hoi et al. [8], 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 main idea of unsupervised distance metric learning methods is to learn low-dimensional manifold where distances between most of observed data are preserved. These methods 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 Eigenamp (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) [12], Multidimensional Scaling (MDS) [13] and the Independent components analysis (ICA) [14].

For semi-supervised 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 method for content-based image retrieval [16].

In this paper, we introduce a new semi-supervised metric 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 in a semi supervised setting, while AQK assumes that labels are already known. The proposed method maps data vectors into a kernel space and learns relevant and irrelevant features' vectors from classification knowledge using class probability distributions. Based on Quasiconformal transformed kernels, the proposed learning process generates for each class a suitable similarity model by accumulating classification knowledge collected over multiple query sessions.

The next section presents the kernel-based similarity model used in this paper. The proposed semi-supervised 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 investigation a nonlinear similarity measurement based on learning relevant and irrelevant features from classification knowledge. The learning process allows the system to compute, for each class a suitable kernel function for similarity measure. Gaussian radial basis function (GRBF) is used as a kernel defined by:

$$k(a,b) = e^{\frac{-\|a-b\|^2}{2\delta^2}} \quad (1)$$

where δ is a scaling parameter and a and b are two vectors in the input space. The distance between a and b is defined as the inner product of $\Phi(a)$ and $\Phi(b)$, where Φ is the function that maps the vectors a and b from an input space χ to a high dimensional feature space F. The inner product between two vectors a and b, $\langle \Phi(a), \Phi(b) \rangle = k(a,b)$, can be considered as a measure of their similarity. Therefore the distance between a and b is defined as:

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

The advantage of the kernel-based similarity is the ability to create a new kernel function derived from the existing similarity model depending on the considered application. Based on Quasiconformal transformed kernels [5] we can modify the similarity measurement (kernel) in order to reduce the distance around relevant features and to expand the distance around irrelevant ones. 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 (explained in 3.3).

3. SEMI-SUPERVISED METRIC LEARNING STRATEGY

The proposed semi-supervised metric learning strategy 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 projected space. In the final step, we learn the similarity model from data and apply the $k-NN$ search.
In the next sub-section, we introduce the three steps method.

3.1 Step 1: Kernel Principal Component Analysis (KPCA)

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

$$C = \frac{1}{N}\sum_{i=1}^{N} \Phi(x_i)\Phi(x_i)^T \quad (4)$$

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

$$\lambda_l V_l = C V_l \quad (5)$$

where V_l and λ_l are respectively the l^{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:

$$\forall\ p = 1,...M, \exists\ \alpha_{p1}, ..., \alpha_{pN} \in \Re\ s.t. V_p = \sum_{j=1}^{N} \alpha_{pj}\Phi(x_j) \quad (6)$$

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

$$K\alpha_p = \lambda_p \alpha_p \quad (7)$$

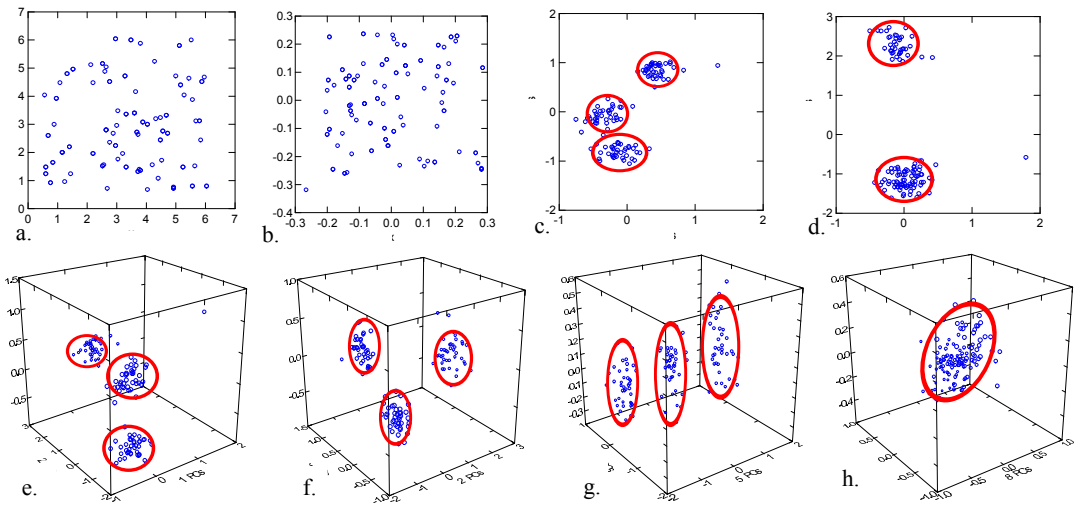


Figure 1. Principal Components for different values of δ and d

where K is the Gram matrix defined by:

$$K_{i,j} = \left\langle \Phi(x_i), \Phi(x_j) \right\rangle$$

where $i, j = 1,...N$ are respectively the row and the column indices of K. The eigenvectors V_l are sorted in a decreasing order according to the magnitudes of their corresponding eigenvalues λ_l, and the first eigenvectors are selected as the basis, that are used to map data vectors into the feature space.

For any input vector x_p, the l^{th} principal component $\tilde{y}_l$ of $\Phi(x_p)$ is given by:

$$\tilde{y}_l = \frac{1}{\sqrt{\lambda_l}} \sum_{j=1}^{N} \alpha_{lj} < \Phi(x_j), \Phi(x_p) > \qquad (8)$$

In KPCA, the nonlinear mapping $\Phi(x)$ into the kernel space is determined by the nonlinear function k (RBF in our case). In the next section, we show that by simply changing the values of δ and d, where δ is the scaling parameter of the GRBF and d is the dimensionality of the feature space, we obtain series of functions which approximate differently the nonlinearity of the original data in the feature space. The optimal kernel function obtained is then used as a similarity measurement.

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 δ 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. To show the influence of the two parameters d and δ on data structure and the classification task, an illustrative example is given in the figure 1.

For this, 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 respectively, from left to right, the values 0.01, 2 and 20 for δ, and a fixed value of $d = 2$. We can see from these figures that the kernel parameter δ has a significant effect on the class separability. When increasing δ until a certain value, better separation of the class vectors is obtained (1.c). In our case, for $\delta = 2$ the best separation is reached, while for a large value of δ (in our case $\delta = 20$), 2 classes are obtained instead of 3 classes.

In the second row, we fixe $\delta = 1$ 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 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 δ values is crucial as it can widely influence the success or the failure of the retrieval systems.

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 δ, which allow to obtain a 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, $m_i = \frac{1}{l_i}\sum_{k=1}^{C_i}\Phi(x_k)$ be the mean vector of class i in the feature space, and $m = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{C_i}\Phi(x_j)$, the mean vector of all vectors N. The average within-cluster distances S_w, which correspond to intra-class variance, and the average between-cluster distances S_b, which correspond to the inter-class variance, can be calculated by:

$$S_b = \frac{1}{N}\sum_{i=1}^{C} l_i (m_i - m)(m_i - m)^T$$

$$S_w = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{C_i}(\Phi(x_j) - m_i)(\Phi(x_j) - m_i)^T$$

The idea is to find the non linear transformation (kernel function which depends on d and δ) that maximizes the following criterion:

$$\gamma = S_b / S_w \tag{9}$$

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. The issue of selecting d eigenvectors, with $d < d'$, d' is the dimensionality of the original data, has been addressed in many works [17]. In our case, we select d eigenvectors in order to capture 98% of the variance, i.e:

$$\rho = \sum_{k=1}^{d}\lambda_k / \sum_{k=1}^{d'}\lambda_k \geq 0.98 \tag{10}$$

To find the values of d and δ 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 δ. An empirical selection is used to determine the initial values of d and δ.

In a first step, we select all the couples whose γ value verifies:

$$\gamma_{d,\delta} \geq 98\%\gamma_{\max} \tag{11}$$

In the second step, we select, among all the couples selected in the first step, those whose statistical variance ρ verifies:

$$\rho_{d,\delta} \geq 98\%\rho_{\max} \tag{12}$$

Finally we keep among the obtained couples those having the lowest value of d and δ. Experimental results (section 4.1) show that the learning strategy used to select the values of (d,δ) achieved the highest performances

3.3 Step 3: learning and searching mechanisms

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 semi-supervised metric learning, we create a new kernel $\tilde{k}$ form the previous one (Equation (3)), and hence a new kernel distance:

$$\begin{aligned} dist(a,b)^2 &= \tilde{k}(a,a) - 2\tilde{k}(a,b) + \tilde{k}(b,b) \\ &= c(a)^2 k(a,a) - 2c(a)c(b)k(a,b) + c(b)^2 k(b,b) \end{aligned} \tag{13}$$

The idea 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. Our aim is to make the space around features farther from or closer to the query, related to their class probability distributions.

$c(x)$ can be computed as follows:

$$c(x) = \frac{P(x/D)}{P(x/S)} \tag{14}$$

where $P(x/D) = \frac{1}{|D|}\sum_{x_i \in D} e^{-\frac{\|x-x_i\|^2}{2\delta^2}}$ and $P(x/S) = \frac{1}{|S|}\sum_{x_i \in S} e^{-\frac{\|x-x_i\|^2}{2\delta^2}}$

$|S|$ 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 classification knowledge, the system identifies $S_i = \{s_{i,1},...s_{i,N}\}$ and $D_i = \{d_{i,1},...d_{i,M}\}$ 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:

$$S = \{S_{i,j}\}_{\substack{i=1,w \\ j=1=Nb}} \quad , \quad D = \{D_{i,j}\}_{\substack{i=1,w \\ j=1=Nb}}$$

Then we compute the suitable similarity model according to equations (13) and (14) for an efficient image retrieval.

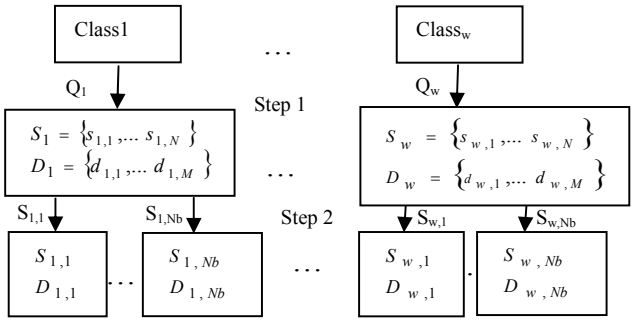


Figure 2. Fonctional diagram of the proposed semi-supervised metric learning.

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 (figure 3). 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 in this section, the first one deals with KPCA parameters estimation strategy, and the second one evaluate our semi-supervised metric learning on the Coil-100 database.

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,δ). 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.

Figure 4.a and 4.b show respectively γ and ρ parameters variations for different values of d (lines) and δ (columns) where darkness corresponds to low values of each parameter. The areas bellow the curves in figures 4.a and 4.b correspond respectively to the values of d and δ which verify equations (11) and (12). The optimal values of d and δ are located around the intersection point of the two curves, illustrated by a circle in figure 4.c. In our tests, the optimal values are $(d = 61, \delta = 11)$.

In the second experiment, we have compared the similarity search quality using the kernel function for different values of the couple (d,δ) (optimal and non optimal). The retrieval quality was also compared with the use of Euclidian distance .The comparison results in term of average recall and precision are given in figure 5. 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 Semi-supervised learning strategy evaluations

In this experiment, we compare the retrieval performances obtained with our semi-supervised 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 uses the pairwise similarity constraints to learn Mahalanobis distance, which consists on assigning large weight to relevant features and low weights to irrelevant ones. It tends to maximize the total intra-class variance and to minimize the total inter-class variance.

To measure the retrieval performance, a set of 600 images are randomly selected from the database and are used as queries. Figure 6 (a) and 6 (b) shows the retrieval results on the Coil-100 database.

We can see that our semi-supervised 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 6 (b) presents the retrieval results, and we can see that our method still outperforms kernel metric, and Mahalanobis metric learning with kernel DCA.

5. CONCLUSION

In this paper, we have proposed an efficient semi-supervised metric learning method to boost the retrieval performance continuously by accumulating classification knowledge collected over multiple query session. Not only

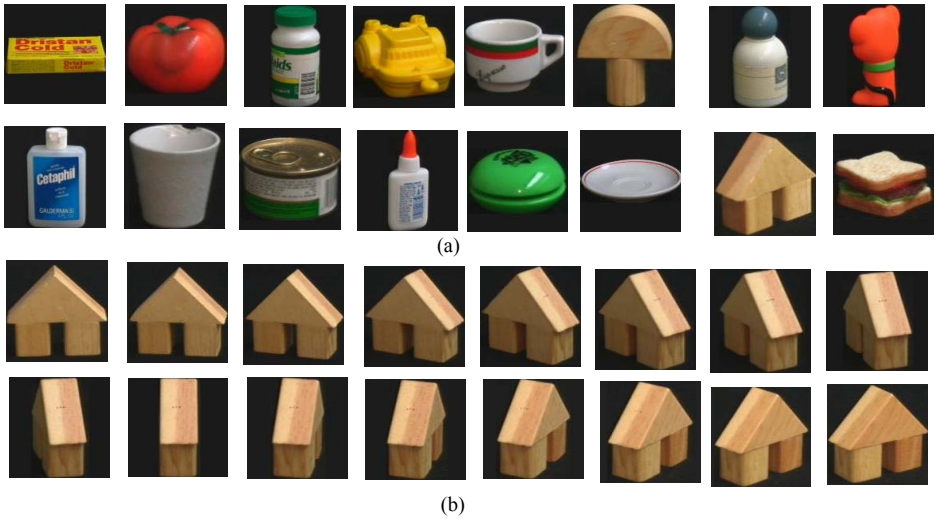


Fig. 3. An example of images (a) and classes (b) of COIL-100 database

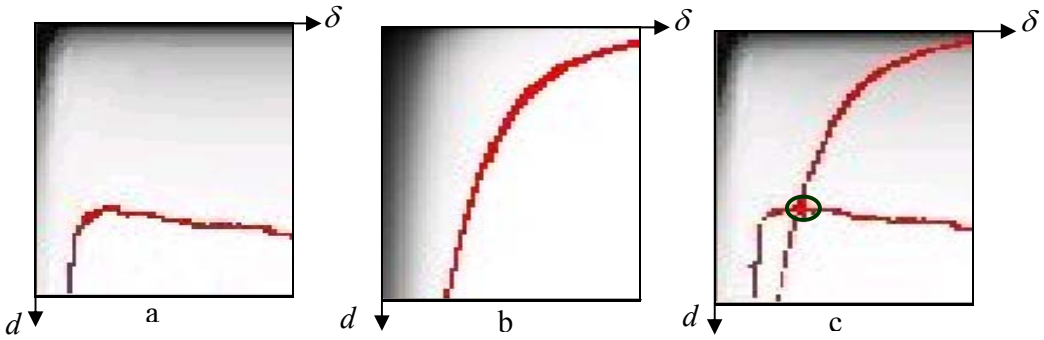


Figure 4. a. $\gamma(d,\delta)$ **b.** $\rho(d,\delta)$ **c.** optimal values of (d,δ)

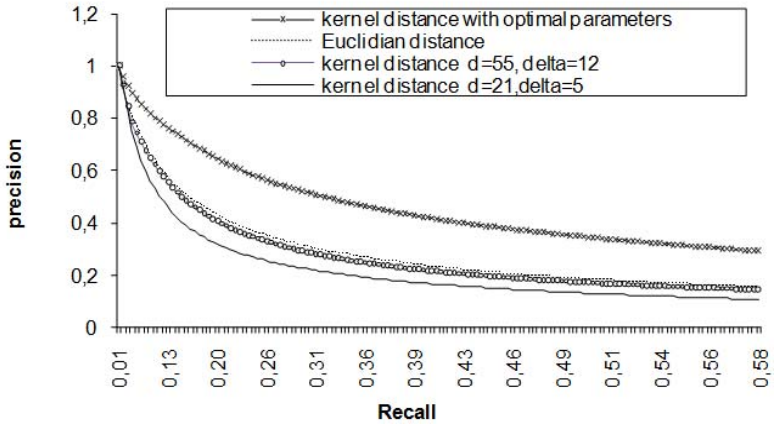
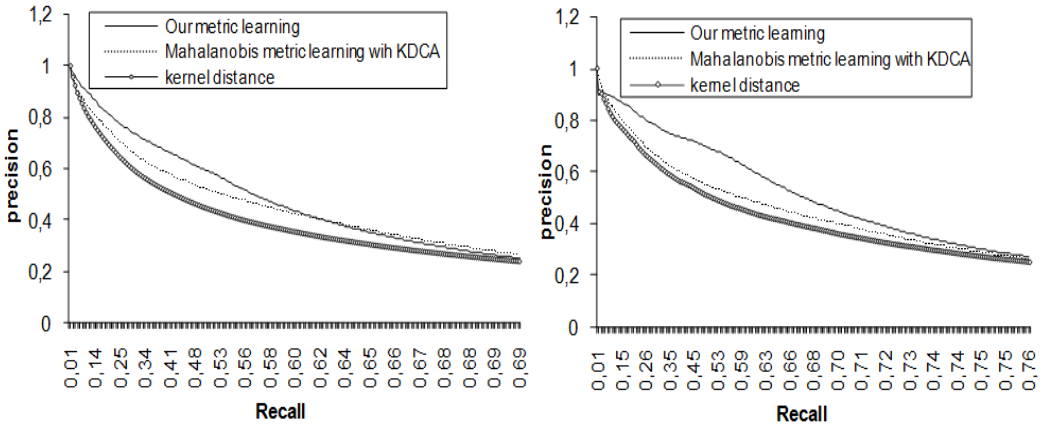


Figure 5. Recall and precision curves under different δ and d values using coil-100 database



(a) (b)

Figure 6. (a) Retrieval results on the coil-100 database. **(b)** Retrieval results on the coil-100 database based on a separate set of query images

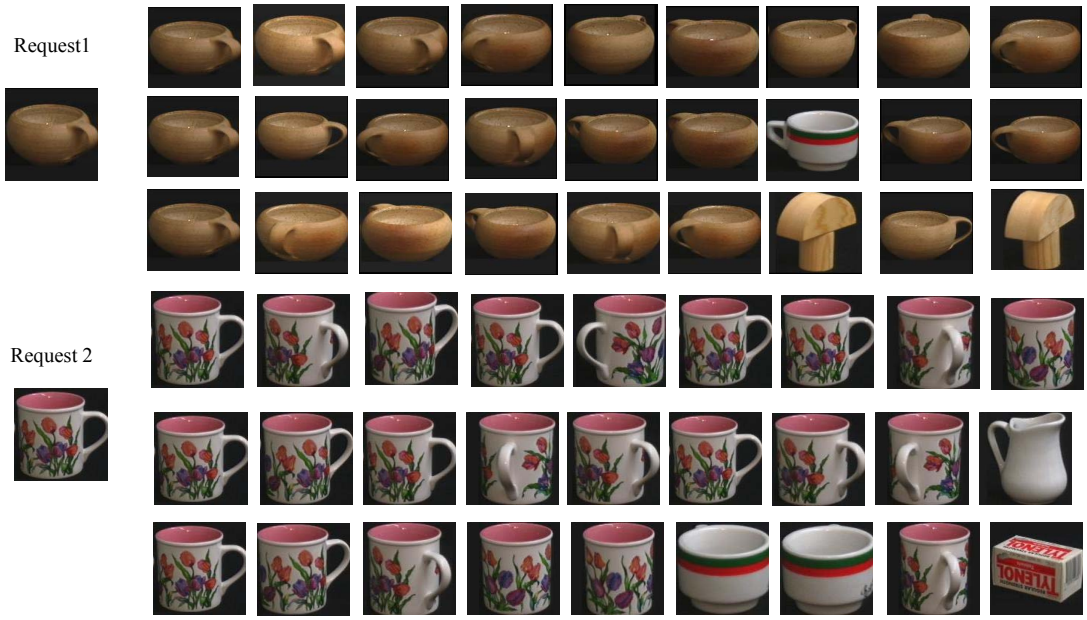


Figure 7. Retrieval results on Coil-100 database using for the 1st row, semi-supervised metric learning, for the 2nd row the Mahalanobis metric learning with KDCA and for the 3rd row kernel distance without learning

does our method based on Quasiconformal kernels improves the retrieval performance of kernel distance, it also outperforms the Mahalanobis metric learning with kernel DCA due to it higher flexibility in metric learning. As future work, we aim to address the main limitation of this approach which resides in the higher computational time. Other possible research direction is to apply the kernel-based metric learning to other pattern recognition tasks.

ACKNOWLEDGEMENT

This work was supported in part by STIC French-Morocco program (Sciences et Technologies de l'Information et de la communication)

6. REFERENCES

[1] M. Flickner. et al., "Query by image and video content: The qbic system", *IEEE Computers*. 1995

[2] E. Xing, A. Ng, M. Jordan, and S. Russell, “Distance metric learning with application to clustering with side-information,” I*n Proc*. NIPS, 2003.

[3] S. Xiang, F. Nie, C. Zhang “ learning a Mahalanobis distance Metric for data clustering and classification”, *In Pattern Recognition* pp. 3600-3612, 14, 2008

[4] L. Yang, “An overview of Distance Metric Learning,” *In Proc. Computer Vision and Pattern recognition*, October 2007.

[5] J. Peng, D. Heisterkamp, and H. Dai, “Adaptive Kernel metric nearest neighbor classification,” *In Proc. International Conference on Pattern Recognition,* 2002.

[6] S. Bermejo and J. Cabestany, “Large margin nearest neighbor classifiers,” I*n Proceedings of the 6th InternationalWork-Conference on Artificial and Natural Neural Networks*. London, UK: Springer-Verlag, pp. 669–676, 2001.

[7] C. Domeniconi, J. Peng, and D. Gunopulos, “Locally adaptive metric nearest-neighbor classificaton,” *IEEE Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1281–1285, 2002.

[8] Hoi, S.C.H.; Wei Liu; Lyu, M.R.; Wei-Ying Ma. “ Learning distance metrics with contextual constraints for image retrieval,” In *Computer Vision and Pattern Recognition*, Volume 2, Page(s): pp. 2072 - 2078, 2006

[9] J.B. Tenenbaum, V. de Silva, and J. C. Langford. “ A global geometric framework for nonlinear dimensionality reduction”. Science, 290, 2000.

[10] S. Roweis and L. Saul. “ Nonlinear dimensionality reduction by locally linear embedding”. In Science, 2000.

[11] M. Belkin and P. Niyogi. “Laplacian eigenmaps for dimensionality reduction and data representation”. Neural Computation, 15(6), 2003.

[12] Schölkopf B., Smola A., Muller K.-R.: “Nonlinear component analysis as a kernel eigenvalue problem, ” *In Neural Computation* 10 , pp. 1299–1319, 1998

[13] T. Cox and M. Cox. “Multidimensional Scaling”. Chapman and Hall, 1994.

[14] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. “Learning a mahalanobis metric from equivalence constraints”. JMLR, 6:937–965, 2005.

[15] P. Common. “Independent component analysis — a new concept?” Signal Processing, 36:287–314, 1994.

[16] W .Y. KIM, Y.S.Kim: “A new region-based shape descriptor,” TR# 15-01, December 1999

[17] A. Pentland, B. Moghaddam, T. Starner, “View-based and modular eigenspace for face recognition,” In Proceedings of the International Conference on Computer Vision, pp. 84–91, 1994.

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

[19] Idrissi, K., lavoué , G., Ricard , J., and Baskurt A.: “ Object of Interest based visual navigation, retrieval and semantic content identification system”, In Computer Vision and Image Understanding , 94(1): pp.271-294, 2004

[20] Manjunath B. S., Salembier P., Sikora T. (Eds.): “ Introduction to MPEG-7”, Wiley, 2002

[21] Hong. C, Dit-Yan. Y, “Kernel distance metric learning for content-based image retrieval” Image and vision computing, volume 25, Issue 5, May 2007