PAPER Special Section on Information-Based Induction Sciences and Machine Learning

Kernel Optimization Based Semi-Supervised KBDA Scheme for Image Retrieval

Xu YANG[†a)], ***Student Member***, **Huilin XIONG**[†], ***and*** **Xin YANG**[†], ***Nonmembers***

SUMMARY Kernel biased discriminant analysis (KBDA), as a subspace learning algorithm, has been an attractive approach for the relevance feedback in content-based image retrieval. Its performance, however, still suffers from the "small sample learning" problem and "kernel learning" problem. Aiming to solve these problems, in this paper, we present a new semi-supervised scheme of KBDA (S-KBDA), in which the projection learning and the "kernel learning" are interweaved into a constrained optimization framework. Specifically, S-KBDA learns a subspace that preserves both the biased discriminant structure among the labeled samples, and the geometric structure among all training samples. In kernel optimization, we directly optimize the kernel matrix, rather than a kernel function, which makes the kernel learning more flexible and appropriate for the retrieval task. To solve the constrained optimization problem, a fast algorithm based on gradient ascent is developed. The image retrieval experiments are given to show the effectiveness of the S-KBDA scheme in comparison with the original KBDA, and the other two state-of-the-art algorithms.

key words: *multimedia information systems, content based image retrieval, biased discriminant analysis, kernel optimization, relevance feedback*

1. Introduction

With the rapid growth of digital image records and the rapid increase of computer power, content-based image retrieval (CBIR) becomes one of the most active research fields [1] in the last decade. Basically, there are two factors affecting the performances of an image retrieval system, the visual features, extracted from images, and the distance metric, used to measure the similarity between two image samples. Given a specific feature representation, the performance of a retrieval system depends heavily on the similarity metric. Instead of using a predetermined metric, more promising approach is to learn an appropriate metric from the returning results of the relevance feedback [2], [3].

In the past years, many subspace learning algorithms, such as LDA [4], ERCA [5], DCA [6], BDA [7], etc. are applied to find a good distance metric, among which the BDA scheme shows the superiority in image retrieval over other methods, because of its "biased" strategy. The "biased" means that BDA only let the relevant samples be clustered closely in the discriminant subspace, but does not apply to the irrelevant samples. Since the irrelevant samples usually come from multiple image classes, this "biased" strategy is more reasonable for the task of image retrieval. Moreover, to capture the nonlinear discriminant components, the kernel trick [8] has been introduced to the BDA scheme to formulate the kernel version of biased discriminant analysis, denoted by KBDA. Many investigations [9]–[11] show that KBDA outperforms SVM [12], KFDA [13], etc. in image retrieval.

However, KBDA has two shortcomings when applied to the relevance feedback scheme in CBIR. First, in practice, the number of the returned images in the relevance feedback scheme is usually limited. Therefore, the learned projection by KBDA may not well adapt to other images of the database. Secondly, kernel selection [14] is crucial for the performance of KBDA. Simple kernel cannot capture the intrinsic nonlinear structure of data, whereas complicated kernels may result in over-fitting. Furthermore, in image retrieval, the training data set is always in change with the user feedback, which makes the situation even worse. In this paper, we refer to these two problems as "small sample learning" problem and "kernel learning" problem, respectively.

To address these problems, we develop a semi-supervised scheme of KBDA learning (S-KBDA), which simultaneously formulates the projection learning and the "kernel learning" into a constrained optimization framework. Specifically, to handle the "small sample learning" problem, we adopt a semi-supervised way in the biased discriminant learning, by adding the neighborhood information of the unlabelled samples. In this way, the S-KBDA scheme can learn not only the biased discriminant structure among the labeled samples, but also the geometric structure among the unlabeled samples. Secondly, instead of learning the parameters of a specific kernel function, we take the kernel matrix as the target to be learned, which makes the kernel learning more flexible. Moreover, to maintain the quick retrieval, our scheme first focuses on learning the labeled part of the kernel matrix, and then estimates the other parts of the kernel matrix using a recently developed technique, referred as Kernel Propagation [15]. The final constrained optimization in our scheme is solved by an efficient gradient-based algorithm.

The paper is organized as follows. In Sect. 2, we first review the KBDA algorithm, and then present our S-KBDA scheme. An efficient optimization algorithm to solve the S-KBDA learning is also given in this Section. Section 3 presents the experimental results on an image retrieval database. Finally, Sect. 4 concludes this paper.

Manuscript received December 30, 2010.
Manuscript revised May 5, 2011.

[†]The authors are with the Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China.
a) E-mail: feilang@foxmail.com

DOI: 10.1587/transinf.E94.D.1901

2. Problem Formulation

2.1 Kernel Biased Discriminant Analysis

Let $\{x_1, x_2, \ldots, x_m\} \in R^d$ be m image feature vectors, in which l vectors are labeled as relevant (positive) samples or irrelevant (negative) samples to the query, and the remained $m - l$ vectors are unlabeled. Without loss of generality, let us assume that the first l_p samples belong to the positive class, and the following l_n samples belong to the negative class, $(l_p + l_n = l)$. Employing the kernel function, each feature vector is mapped to a high dimensional kernel space F, $x_i \rightarrow \phi(x_i) \in F$. Essentially, the KBDA scheme tries to find an optimal linear projection from F to a lower dimensional subspace S, so that the positive samples would be well clustering together and the negative ones be pushed away from the positive ones as far as possible. This optimal KBDA projection can be obtained by maximizing the following objective function [7]:

$$J(w) = \frac{w^T S_{np} w}{w^T S_p w}$$
$$S_p = \sum_{i=1}^{l_p} [\phi(x_i) - m_p][\phi(x_i) - m_p]^T$$
$$S_{np} = \sum_{i=l_p+1}^{l} [\phi(x_i) - m_p][\phi(x_i) - m_p]^T \tag{1}$$

where $m_p = \frac{1}{l_p} \sum_{j=1}^{l_p} \phi(x_j)$ is the mean vector of the labeled positive samples in F. We call S_p the positive scatter matrix and S_{np} the negative biased scatter matrix.

Let Φ denote the mapped data matrix in F, which contains the labeled samples $\Phi_l = (\phi(x_1), \ldots, \phi(x_l))$ and the unlabeled samples $\Phi_u = (\phi(x_{l+1}), \ldots, \phi(x_m))$. We can rewrite the scatter matrices as follows

$$S_p = \Phi_l \left[\sum_{i=1}^{l_p} \left(\mathbf{e}_i - \frac{1}{l_p} \sum_{j=1}^{l_p} \mathbf{e}_j \right) \left(\mathbf{e}_i - \frac{1}{l_p} \sum_{j=1}^{l_p} \mathbf{e}_j \right)^T \right] \Phi_l^T$$
$$= \Phi_l M_p \Phi_l^T$$
$$S_{np} = \Phi_l \left[\sum_{i=l_p+1}^{l} \left(\mathbf{e}_i - \frac{1}{l_p} \sum_{j=1}^{l_p} \mathbf{e}_j \right) \left(\mathbf{e}_i - \frac{1}{l_p} \sum_{j=1}^{l_p} \mathbf{e}_j \right)^T \right] \Phi_l^T$$
$$= \Phi_l M_{np} \Phi_l^T \tag{2}$$

where e_j is an l-dimensional unit vector whose entries are all 0 except the j-th, which is 1. M_p and M_{np} are two $l \times l$ constant matrices, which only depend on the label of the data.

By the representer theorem [16], we know that the solution to the learning problem in (1) can be expressed as

$$w = \Phi_l \beta \tag{3}$$

where $\beta \in R^l$ is the coefficient vector. Therefore, optimizing the projection vector w means to find the optimal coefficient vector β. Since the objective function $J(w)$ is invariant with respect to rescaling of the projection vector w, we can always choose β such that the denominator in problem (1) equals to 1. Therefore, the problem of maximizing $J(w)$ is transformed to a constrained optimization problem. Specifically, substituting (2) and (3) into (1), the constrained optimization problem can be formulated as

$$\max_{\beta} \beta^T K_{ll} M_{np} K_{ll} \beta$$
$$\text{subject to } \beta^T K_{ll} M_p K_{ll} \beta = 1 \tag{4}$$

where K_{ll} is the labeled part of the kernel matrix $K = \Phi^T \Phi = \begin{pmatrix} K_{ll} & K_{lu} \\ K_{lu}^T & K_{uu} \end{pmatrix}$.

The columns of the optimal β are the eigenvectors of $\left(K_{ll} M_p K_{ll} + \mu I\right)^{-1} K_{ll} M_{np} K_{ll}$ corresponding to the non-zero eigenvalues, where μ is a regularization factor, set empirically to 0.01 in our experiment. Thus, the projection of sample x_j in the KBDA subspace is

$$f\left(x_j\right) = w^T \phi\left(x_j\right) = \beta^T \begin{pmatrix} k_{1j} \\ \vdots \\ k_{lj} \end{pmatrix} \tag{5}$$

where $f(\cdot)$ is called the projection function.

2.2 Semi-Supervised Kernel Biased Discriminant Analysis

KBDA learns the optimal discriminant projection only from the labeled samples. However, in the case of image retrieval, where the number of labeled samples from the relevance feedback is quite small, the subspace learned by KBDA may not be optimal to other unlabeled samples. This could lead to serious over-fitting. An effective way to improve the robustness of the KBDA learning is to utilize the intrinsic structure information provided by the abundant unlabeled samples. To do so, we introduce the smoothness regularization technique, according to the spectral graph theory [17], to the KBDA scheme, aiming to preserve the local neighboring relation of the training samples.

Let $G = (V, S)$ be an undirected, weighted graph constructed on the whole dataset, with the node set $V = \{x_i\}_{i=1}^m$ and the weight matrix $S = \left[s_{ij}\right]_{m \times m}$. s_{ij} measures the similarity between nodes x_i and x_j, which can be calculated using various similarity criteria, such as the local neighborhood relationship as in [18], the heat kernel similarity as in [19], and the binary similarity frequently used in [20]–[22]. For the sake of computational efficiency, the simple-minded binary similarity is used in this paper, which is defined as follows:

$$s_{ij} = \begin{cases} 1, \text{ if } x_i \in N_p\left(x_j\right) \text{ or } x_j \in N_p\left(x_i\right) \\ 0, \text{ otherwise.} \end{cases} \tag{6}$$

where $N_p(x_i)$ denotes the set of p-nearest neighbors of x_i. The so called normalized graph Laplacian is defined as $L = D^{-\frac{1}{2}}(D - S) D^{-\frac{1}{2}}$, where D is a $m \times m$ diagonal matrix with $d_{ii} = \sum_{j=1}^m s_{ij}$. Following the idea of regularization in spectral graph theory [23], we can measure the smoothness of the KBDA projection function $f(\cdot)$ on dataset $\{x_i\}_{i=1}^m$ by

$$S(f) = \frac{1}{2}\sum_{i,j=1}^{m} s_{ij}\left\|\frac{f(x_i)}{\sqrt{d_{ii}}} - \frac{f(x_j)}{\sqrt{d_{jj}}}\right\|^2 = f_m^T L f_m \tag{7}$$

where $f_m = (f(x_1), \cdots, f(x_m))^T$. We can see that the value of $S(f)$ "penalizes" the large change of the KBDA embeddings between two points linked with a large weight. In other words, minimizing $S(f)$ is in accordance with the goal of preserving the neighboring relationship. From (5) and (7), we have

$$S(f) = \beta^T (K_{ll} \;\; K_{lu}) L (K_{ll} \;\; K_{lu})^T \beta \tag{8}$$

Adding $S(f)$ into (4) as a regularizer, we got the semi-supervised KBDA objective function.

$$\begin{aligned}&\max_{\beta} \beta^T \left[K_{ll} M_{np} K_{ll} - \alpha (K_{ll} \;\; K_{lu}) L (K_{ll} \;\; K_{lu})^T\right]\beta\\ &\text{subject to } \beta^T K_{ll} M_p K_{ll} \beta = 1\end{aligned} \tag{9}$$

where α is a suitable constant, and $0 \le \alpha \le 1$. In our experiments, α is empirically set to be 0.5.

2.3 S-KBDA with Kernel Learning

Like other kernel methods, the performance of KBDA or S-KBDA is impacted seriously by the kernel selection. To handle the kernel selection, Zhou et al. [24] proposed to use the kernel partial alignment, and Wang [10] proposed to use the trace ratio between the scatter matrices to measure the goodness of a kernel. Both of them are supervised approaches aiming to choose a good spread parameter δ for the Gaussian kernel. However, just like the subspace learning, the kernel learning in the KBDA scheme is subject to a small number of labeled samples, which still makes the kernel learning suffer from the "small sample learning" problem. Besides, the Gaussian kernel does not change the rank of similarities between retrieval images and the query image, since all the data are mapped into the kernel space in a uniform way. Therefore, the distribution of the data cannot be optimized effectively, and the improved retrieval performance using the optimal Gaussian kernel is very limited.

Instead of learning a kernel function, our S-KBDA scheme learn the kernel matrix K. From (9) we see that the S-KBDA learning is only influenced by the sub-kernel matrix K_{ll} and K_{lu}. To accelerate the computation of our algorithm, we only focus our optimization on K_{ll}. Once K_{ll} is learned, K_{lu} can be obtained efficiently by using a technique, called Kernel Propagation (KP) [15]. KP aims to propagate the learned small sub-kernel matrix into a large-sized full-kernel matrix. This propagation is based on the consistency assumption of the kernel map $\phi(\cdot)$ (for more details, please see [15]). According to the KP algorithm, K_{lu} can be calculated from K_{ll} as follows

$$K_{lu} = -K_{ll} L_{lu} L_{uu}^{-1} \tag{10}$$

where L_{lu} and L_{uu} are the corresponding sub-matrices of the graph Laplacian $L = \begin{pmatrix} L_{ll} & L_{lu} \\ L_{lu}^T & L_{uu}\end{pmatrix}$.

From (10) and (9), keeping in mind that K_{ll} also needs to be optimized for kernel learning, we formulate our optimization problem as:

$$\begin{aligned}&\max_{K_{ll},\beta} \beta^T K_{ll}\left(M_{np} - \alpha\left(L_{ll} - L_{lu} L_{uu}^{-1} L_{lu}^T\right)\right) K_{ll}\beta\\ &\text{subject to } \beta^T K_{ll} M_p K_{ll} \beta = 1\\ &K_{ll} \ge 0\end{aligned} \tag{11}$$

2.4 The Optimization Algorithm

It is quite difficult to obtain the global optimum of the above problem directly. Thus, we introduce the optimal value function [25] $M(K_{ll})$ for problem (11) by keeping K_{ll} fixed.

$$\begin{aligned}&\mathcal{M}(K_{ll}) := \max_{\beta} \beta^T K_{ll} \tilde{M}_{np} K_{ll} \beta\\ &\text{subject to } \beta^T K_{ll} M_p K_{ll} \beta = 1\end{aligned} \tag{12}$$

where $\tilde{M}_{np} = M_{np} - \alpha\left(L_{ll} - L_{lu} L_{uu}^{-1} L_{lu}^T\right)$ is a constant $l \times l$ matrix. The global (or local) maximum of $\mathcal{M}(K_{ll})$ can be viewed as a global (or local) maximum of problem (11). Therefore, instead of solve problem (11) directly, we employ a gradient-based algorithm to maximize $\mathcal{M}(K_{ll})$. To calculate the derivative of the optimal value function $\mathcal{M}(K_{ll})$, let us first present the following theorem.

Theorem 1: The derivative of $\mathcal{M}(K_{ll})$ can be calculated as:

$$\frac{\partial \mathcal{M}(K_{ll})}{\partial k_{ij}} = \bar{\beta}^T \frac{\partial K_{ll} \tilde{M}_{np} K_{ll}}{\partial k_{ij}} \bar{\beta} + \bar{\lambda} \bar{\beta}^T \frac{\partial K_{ll} M_p K_{ll}}{\partial k_{ij}} \bar{\beta} \tag{13}$$

where $\bar{\beta}$ is the unique optimal solution of problem (12), $\bar{\lambda}$ is the Lagrange multiplier associated with the equality constraint.

Proof: By using the Lagrange multiplier method, we can change problem (12) to the following unconstrained optimization problem:

$$\mathcal{L}(\beta, \lambda) = \beta^T K_{ll} \tilde{M}_{np} K_{ll} \beta + \lambda\left(\beta^T K_{ll} \tilde{M}_p K_{ll} \beta - 1\right) \tag{14}$$

Therefore, we have $\mathcal{M}(K_{ll}) = \max_{\beta,\lambda} \mathcal{L}(\beta, \lambda) = \mathcal{L}\left(\bar{\beta}, \bar{\lambda}\right)$.

At the maximum point of $\mathcal{L}(\beta, \lambda)$, the following equalities hold:

$$\frac{\partial \mathcal{L}(\beta, \lambda)}{\partial \beta} = 2K_{ll} \tilde{M}_{np} K_{ll} \beta + 2\lambda K_{ll} M_p K_{ll} \beta = \mathbf{0} \tag{15}$$

$$\frac{\partial \mathcal{L}(\beta, \lambda)}{\partial \lambda} = \beta^T K_{ll} M_p K_{ll} \beta - 1 = 0 \tag{16}$$

From (15), we know that β is an eigenvector of the matrix $\left(K_{ll} M_p K_{ll} + \mu I\right)^{-1} K_{ll} \tilde{M}_{np} K_{ll}$ and $-\lambda$ is the corresponding eigenvalue. Left-multiplying Eq. (15) with β^T and taking (16) into it, we have $-\lambda = \beta^T K_{ll} \tilde{M}_{np} K_{ll} \beta$. Considering that

$\beta^T K_{ll}\tilde{M}_{np}K_{ll}\beta$ is the objective function in (12), we know the optimal Lagrange multiplier $\bar{\lambda}$ equals to the negative of the largest eigenvalue of $\left(K_{ll}M_pK_{ll}\right)^{-1}K_{ll}\tilde{M}_{np}K_{ll}$, and the optimal solution $\bar{\beta}$ should be the corresponding eigenvector multiplied by an appropriate scale factor such that Eq. (16) is satisfied.

According to the Theorem 4.1 in [25], since both $\bar{\beta}$ and $\bar{\lambda}$ are unique, $\mathcal{M}(K_{ll})$ is continuously differentiable, its derivative can be calculated as follows

$$\begin{aligned}\frac{\partial \mathcal{L}\left(\bar{\beta},\bar{\lambda}\right)}{\partial k_{ij}} &= \bar{\beta}^T\frac{\partial K_{ll}\tilde{M}_{np}K_{ll}}{\partial k_{ij}}\bar{\beta} + \bar{\lambda}\bar{\beta}^T\frac{\partial K_{ll}M_pK_{ll}}{\partial k_{ij}}\bar{\beta}\\ &\quad + 2\frac{\partial \bar{\beta}^T}{\partial k_{ij}}\left(K_{ll}\tilde{M}_{np}K_{ll}\bar{\beta} + \bar{\lambda}K_{ll}M_pK_{ll}\bar{\beta}\right) \qquad (17)\end{aligned}$$

where the last term equals to 0 because of (15). Hence, the Eq. (13) is established. ■

Let us denote the gradient matrix of $\mathcal{M}(K_{ll})$ by $\nabla\mathcal{M}(K_{ll})$. If we maximize $\mathcal{M}(K_{ll})$ using the general gradient ascent method

$$K_{ll(t)} = K_{ll(t-1)} + \eta\nabla\mathcal{M}(K_{ll}) \qquad (18)$$

where $\eta > 0$ is the learning rate, the positive semidefinite property of K_{ll} may be destroyed. Therefore, we take another search direction δK_{ll}, along which the positive semidefinite property of K_{ll} can be preserved, and at the same time coincide mostly with the steepest ascent direction. This direction can be obtained by solving the following problem

$$\begin{aligned}&\min_{\delta K_{ll}} \|\nabla\mathcal{M}(K_{ll}) - \delta K_{ll}\|_F^2\\ &\text{subject to } \ \delta K_{ll} \geq 0 \qquad (19)\end{aligned}$$

Moreover, according to the Wielandt-Hoffman theorem [26], the solution of (19) can be expressed as $\delta K_{ll} = U\tilde{\Lambda}U$, where $\tilde{\Lambda} = \mathrm{diag}\left(\max\left(0,\lambda_1\right),\cdots,\max\left(0,\lambda_l\right)\right)$, $U = (u_1,\cdots,u_l)$. λ_i and u_i are the eigenvalues and eigenvectors of the non positive semidefinite matrix $\nabla\mathcal{M}(K_{ll})$. To ensure the convergence of the gradient algorithm, a gradually decreasing learning rate is adopted.

$$\eta(t) = \eta_0\left(1 - \frac{t}{N}\right) \qquad (20)$$

where η_0 is the initial learning rate, N denotes the maximum number of iterations, and t represents the current iteration number.

Finally, we summarize the overall optimization procedures in **Algorithm 1**. The algorithm maximizes the S-KBDA objective function iteratively, and in each iteration the two variables β and K_{ll} are optimized alternatively. Specifically, in each iteration of the algorithm, the local optimal coefficient vector β is calculated directly by a generalized eigenvalue decomposition method, and the local optimal sub-kernel matrix K_{ll} is calculated by a gradient updating procedure shown in **Algorithm 2**.

Algorithm 1 The proposed optimization algorithm

```
Input:
    The maximum number of iterations $J$ (e.g., 10), and the preset tolerance value $\varepsilon$ (e.g., $10^{-5}$).
1: Initialize: $K_{ll(0)}$ using the Gaussian kernel;
2: For $i = 1, 2, \cdots, J$ do
    • Compute the projection vector $\beta_i$ and Lagrange multiplier $\lambda_i$ by solving the problem
          $(\beta_i, \lambda_i) = \underset{\beta,\lambda}{\mathrm{argmax}}\, \mathcal{L}(\beta, \lambda)$
    • Compute $K_{ll(i)}$ by solving the problem
          $K_{ll(i)} = \underset{K_{ll}}{\mathrm{argmax}}\, \mathcal{M}(K_{ll})$
    • if $i > 1$ and $|\mathcal{L}(\beta_i, \lambda_i) - \mathcal{L}(\beta_{i-1}, \lambda_{i-1})| < \varepsilon$ and $\left|\mathcal{M}\left(K_{ll(i)}\right) - \mathcal{M}\left(K_{ll(i-1)}\right)\right| < \varepsilon$, then break (converged);
3: Output: The final kernel matrix $K_{ll(i)}$ .
```

Algorithm 2 Compute $K_{ll(i)}$ in step 2 of the Algorithm 1

```
Input:
    The starting matrix $X_0 = K_{ll(i-1)}$;
    The initial learning rate $\eta_0$ (e.g., 0.01), and the maximum number of iterations $N$ (e.g., 10).
1: Initialize: the optimal projection vector $\bar{\beta}$ and the Lagrange multiplier $\bar{\lambda}$ to be $\beta_i$ and $\lambda_i$;
2: For $t = 1, 2, \cdots, N$ do
    • Compute the gradient matrix $\nabla\mathcal{M}(K_{ll})$ at $K_{ll} = X_{t-1}$ by Theorem 1
    • Compute a positive semidefinite matrix $\delta K_{ll}$ to approximate $\nabla\mathcal{M}(K_{ll})$
    • Update the kernel matrix by
          $X_t = X_{t-1} + \eta(t-1)\delta K_{ll}$
    • Decrease the learning rate by $\eta(t) = \eta_0\left(1 - \frac{t}{N}\right)$
3: Output: Set $K_{ll(i)} = X_t$.
```

3. CBIR Experimental Results

3.1 Experimental Setting for Comparison Study

The experimental database that we used consists of 50 categories, each containing exactly 100 images. These images are selected from the COREL photo dataset according to their semantic relevance, such as butterfly, dog, cat and rose, etc. Figure 1 shows some image examples in this dataset. Three types of visual features, that is color, edge and texture, are used to represent the images. (1) Color feature extraction: For each image, color mean, color variance and color skewness are extracted from the H, S and V channel, respectively, to form a 9-dimensional color moment. (2) Edge feature extraction: A total of 36-dimensional edge direction histogram features are calculated by using the Canny edge detector, each direction covering 10 degrees. (3) Texture feature extraction: We apply the Daubechies wavelet transform to each of the gray images to derive a 3-level

Fig. 1 Some samples from the image retrieval dataset.

image decomposition, and then 18-dimensional texture features are extracted by calculating the first two moments of coefficients from the 9 high frequency sub-bands in all the three levels [27]. Totally, we use a 63-dimensional feature vector to represent each image.

For the fairness of comparison, an automatic feedback scheme is used to simulate the real relevant feedback. In the first round of retrieval, the system ranks the images according to their Euclidean distance to the query image. Then, the first five relevant images and five irrelevant images are selected as the positive and negative feedbacks according to their ground truth. With these feedbacks, the learning algorithms are trained to re-rank the image database. In the next round of retrieval, another 10 selections will be used together with the previous ones for training. Note that the selected feedback images in previous training are excluded from present selections.

In the experiment, different algorithms are compared with two procedures. In procedure A, we perform the relevance feedback 4 times for each algorithm, which means the same size of labeled training sets are used, and then the retrieval results of different approaches are compared. Procedure B is designed aiming to evaluate the efficiency of the retrieval systems, where the same target retrieval precision is set in advance, and the number of the labeled images needed to achieve this target is compared. For both procedures, five-fold cross validation is employed to evaluate the average performances of different methods. Specifically, we divide the whole image database into five subsets with equal size. At each trial, one subset is used as the query set, and the other 4 subsets are used as the database for retrieval.

The retrieval performance of the proposed algorithm (S-KBDA) is compared with KBDA, KMMP [22] and SVM. KBDA is what our S-KBDA algorithm is fundamentally based on and one of the best supervised learning algorithms applied in image retrieval [10], [11]. KMMP is a semi-supervised manifold learning algorithm designed for maximizing the margin between positive and negative examples at each local neighborhood. The KMMP algorithm has been shown to get the best performance with two embedding dimensions in [22]. As for S-KBDA and KBDA, the best retrieval performance is obtained with 2 to 5 dimensions for different image categories in our experiments. But their differences in average retrieval precision of all categories are negligible. Therefore, we just also set the embedding dimension to be 2 for its less computation cost. After being projected into the 2-dimensional subspace by S-KBDA, KBDA or KMMP, the database are re-ranked by their Euclidean distances to the query. SVM is a supervised learning algorithm, which learns the hyperplane to separate the positive and negative samples with a maximal margin. The signed distances to this hyperplane are used to re-rank the database [12].

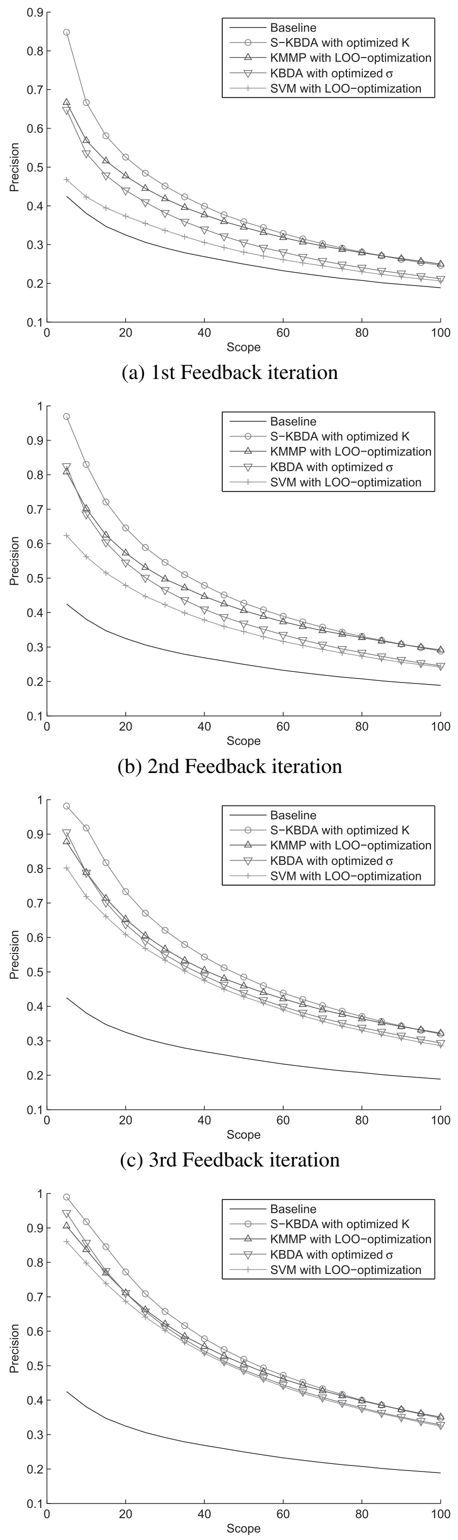
In the experiment, the Gaussian kernel is used to construct the kernel matrix. For KBDA and S-KBDA, since the kernel can be optimized automatically during the learning process, we just set the initial kernel parameter δ to be 1. The kernel parameter δ for KMMP, the kernel parameter δ and the regularization parameter C for SVM are both selected by using the Leave-One-Out cross validation. The graph Laplacian matrix used in S-KBDA and KMMP is constructed from the 5-nearest neighbors, and the unsupervised dataset is composed of the top 400 images ranked by the Euclidean distance.

To get a fast retrieval response, the maximum number of iterations in **Algorithm 1** and **Algorithm 2**, J and N, should be limited, and accordingly the initial learning rate η_0 should be large enough to guarantee that the learned parameters are close to their optimal values. In the experiments, we set them empirically to be 10, 10 and 0.01, respectively, based on the following observations: About 75% of the optimizing procedures could be converged in less than 10 iterations. Using large values of J and N, does not necessarily lead to a remarkable performance improvement in retrieval precision, but does lead to a more complicated computation, which can make the retrieval system inefficient.

3.2 Results and Discussions in Procedure A

Figure 2 shows the precision-scope curves of different algorithms after each round of feedbacks. The baseline curve describes the initial retrieval result by using the Euclidean distances in the original 63-dimensional space. Based on the results shown in Fig. 2, we observe that both S-KBDA and KMMP consistently outperform the supervised learning algorithms of KBDA and SVM, which indicates that the unlabeled images are very helpful to improve the retrieval performance. This improvement is especially obvious in the first two rounds of feedbacks where the number of labeled images is very small.

For the two semi-supervised learning methods, S-KBDA and KMMP, we see that S-KBDA performs better than KMMP in all the feedback iterations. This may be caused by the following two reasons. One is that the biased strategy in S-KBDA only pushes the positive samples close together and does not do this to the negative ones, while KMMP push all the samples with the same label close. The biased strategy is especially helpful in the case when we lack negative samples. The superiority of the biased strategy over the balanced strategy can also be verified by comparing the performances between SVM and KBDA. The other reason



(a) 1st Feedback iteration

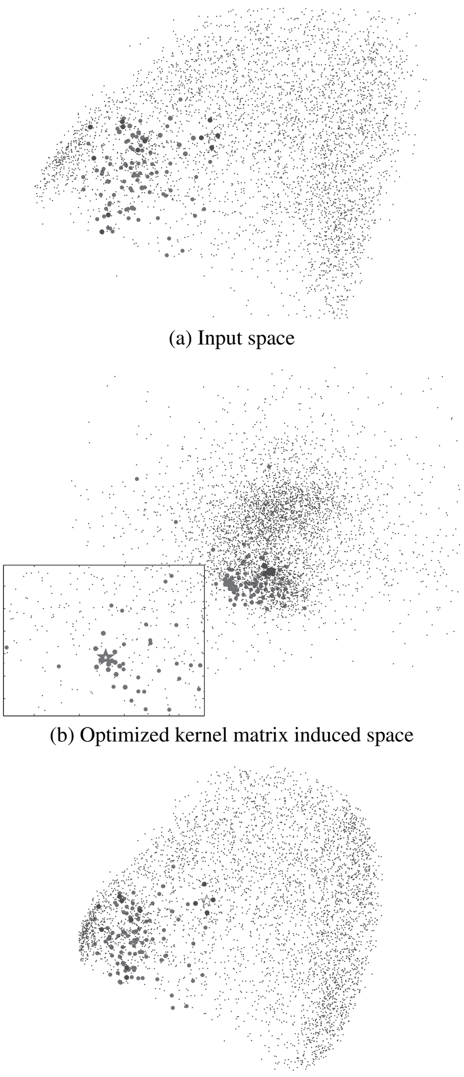
(b) 2nd Feedback iteration

(c) 3rd Feedback iteration

(4) 4th Feedback iteration

Fig. 2 The precision-scope curves for all the four feedback iterations.

lies in the fact that different kernels are used in S-KBDA and KMMP to enhance the nonlinear capability. The Gaussian kernel with an optimal parameter is used in KMMP, whereas S-KBDA learns the kernel itself in a nonparametric manner.



(a) Input space

(b) Optimized kernel matrix induced space

(c) Optimal Gaussian kernel induced space

Fig. 3 2-dimensional semantic visualization.

Since the data is expected to have better linear separability in the kernel space, we conduct a visualization experiment to compare these two kernels. The results are shown in Fig. 3, where the star denotes the query image, the bigger (smaller) points represent the labeled (unlabeled) relevant and irrelevant images, respectively. Figure 3 (a) illustrates the original data distribution, where all the data are projected onto its first two PCA directions. Figure 3 (b) and Fig. 3 (c) show the distribution of data in the kernel space by projecting all the data onto its first two KPCA directions. From Fig. 3 (b), we can see that the order of distances between retrieval images and the query image is improved remarkably by the optimized kernel matrix, and therefore, the distribution of samples in this space is more suited to the task of retrieval. However, this improvement is very limited for the Gaussian kernel, see Fig. 3 (c), since it maps the samples in a uniform way.

The average computational time of different algorithms for processing one user's query is given in Table 1. All these

Table 1 Average runtime for processing one query.

	Time at different feedback iterations (s)			
	1	2	3	4
SVM	0.185	0.231	0.267	0.296
KBDA	0.217	0.294	0.339	0.386
KMMP	0.288	0.360	0.402	0.454
S-KBDA	0.306	0.375	0.430	0.477

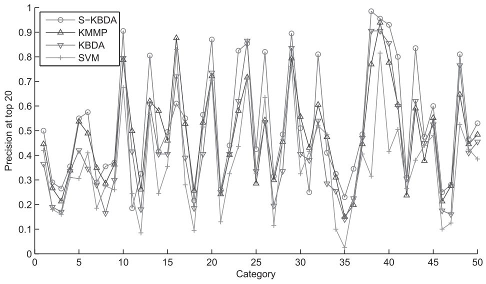


Fig. 4 Precision at the top 20 returns after the 1st feedback iteration.

four algorithms can respond to the query very fast, that is, within 0.5 s. SVM is the fastest and KBDA is slightly slower than SVM. Our S-KBDA is as fast as KMMP and is slower than the supervised methods of SVM and KBDA. All of the experiments were performed on a Pentium IV 3.20-GHz Windows XP machine with a 2-Gbyte memory.

3.3 Results and Discussions in Procedure B

In a real image retrieval system, the returned images on the first screen shot (for example, 20 images in our experiment) are very important to the user. Figure 4 shows the top-20 retrieval precision for different categories after the first feedback iteration. Among all the 50 categories, our S-KBDA approach performs best on 41 categories. For the remaining 9 categories, KMMP performs best on 7 of them, KBDA and SVM each performs best on only one category. We can also see the retrieval performances of these algorithms vary with different categories. There are 14 relatively easy categories on which the retrieval precisions of S-KBDA are higher than 80%. Therefore, for these categories, if the target retrieval precision is set to be 80%, there is no need for the user to teach the system further.

Since it is difficult to require the user go through many rounds of feedbacks, the labor cost to get a satisfied retrieval result is also a critical factor to evaluate a retrieval system. We measure this cost by using the total number of feedback images used in training the system to achieve a target retrieval precision. Figure 5 shows the average number of the labeled images used by the four algorithms to achieve the retrieval precision of 50%, 60%, 70% and 80%, respectively, on the first screen shot. We can see from Fig. 5 that S-KBDA can always achieve the target precision with fewer feedback images. This could save the user a lot of labor and at the same time make the retrieval task finished as soon as possible.

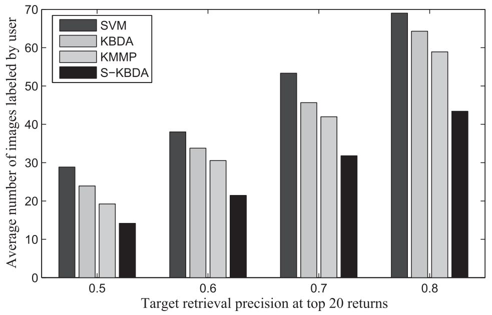


Fig. 5 Average labor cost to achieve the target retrieval precision.

4. Conclusions

This paper presents a new semi-supervised subspace learning algorithm, called S-KBDA, in which the projection vector learning and the “kernel learning” are interweaved into a constrained optimization framework. In the S-KBDA subspace, both the biased discriminant structure and the intrinsic geometric structure can be well preserved, which relieves the over-fitting problem caused by lacking of labeled samples. Moreover, in the procedure of optimization, we use the kernel matrix as the target to be optimized, which makes the learned kernel more flexible and appropriate for the retrieval task. In comparison with the other three state-of-the-art algorithms, that is, KBDA, KMMP and SVM, the retrieval experiments on Corel photo database demonstrate the effectiveness of the proposed scheme.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.60775008), and the National High Technology Research and Development Program (863 Program) of China (Grant No.2007AA01Z196).

References

[1] Y. Liu, D. Zhang, G. Lu, and W.Y. Ma, “A survey of content-based image retrieval with high-level semantics,” Pattern Recognit., vol.40, no.1, pp.262–282, 2007.

[2] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra, “Relevance feedback: A power tool for interactive content-based image retrieval,” IEEE Trans. Circuits Syst. Video Technol., vol.8, no.5, pp.644–655, 1998.

[3] X.S. Zhou and T.S. Huang, “Relevance feedback in image retrieval: A comprehensive review,” Multimedia Syst., vol.8, no.6, pp.536–544, 2003.

[4] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.

[5] D.Y. Yeung and H. Chang, “Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints,” Pattern Recognit., vol.39, no.5, pp.1007–1010, 2006.

[6] C.H. Hoi, W. Liu, M.R. Lyu, and W.Y. Ma, “Learning distance metrics with contextual constraints for image retrieval,” IEEE Conference on Computer Vision and Pattern Recognition, pp.2072–2078, vol.2, New York, USA, 2006.

[7] X.S. Zhou and T.S. Huang, "Small sample learning during multimedia retrieval using biasmap," IEEE Conference on Computer Vision and Pattern Recognition, pp.11–17, 2001.

[8] B. Schölkopf and A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, The MIT Press, 2002.

[9] X.S. Zhou and T.S. Huang, "Comparing discriminating transformations and SVM for learning during multimedia retrieval," 9th ACM International Conference on Multimedia, pp.137–146, Ottawa, Canada, 2001.

[10] L. Wang, K.L. Chan, and P. Xue, "A criterion for optimizing kernel parameters in KBDA for image retrieval," IEEE Trans. Syst. Man Cybern. B, Cybern. vol.35, no.3, pp.556–562, 2005.

[11] D. Tao, X. Tang, X. Li, and Y. Rui, "Direct kernel biased discriminant analysis: A new content-based image retrieval relevance feedback algorithm," IEEE Trans. Multimed., vol.8, no.4, pp.716–727, 2006.

[12] G. Guo, H. Zhang, and S. Li, "Distance-from-boundary as a metric for texture image retrieval," International Conference on Acoustics, Speech, Signal Processing, pp.1629–1632, 2001.

[13] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K.R. Mullers, "Fisher discriminant analysis with kernels," IEEE Workshop on Neural Networks for Signal Processing, pp.41–48, 1999.

[14] H. Xiong, M.N.S. Swamy, and M.O. Ahmad, "Optimizing the kernel in the empirical feature space," IEEE Trans. Neural Netw., vol.16, no.2, pp.460–474, 2005.

[15] E. Hu, S. Chen, D. Zhang, and X. Yin, "Semisupervised kernel matrix learning by kernel propagation," IEEE Trans. Neural Netw., vol.21, no.11, pp.1831–1841, 2010.

[16] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, and K.R. Muller, "Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces," IEEE Trans. Pattern Anal. Mach. Intell., vol.25, no.5, pp.623–628, 2003.

[17] F.R.K. Chung, Spectral Graph Theory, American Mathematical Society Providence, 1997.

[18] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol.290, no.5500, pp.2323–2326, 2000.

[19] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," Neural Comput., vol.15, no.6, pp.1373–1396, 2003.

[20] X. He and P. Niyogi, "Locality preserving projections," in Advances in Neural Information Processing Systems, pp.153–160, MIT Press, 2003.

[21] M.S. Baghshah and S.B. Shouraki, "Kernel-based metric learning for semi-supervised clustering," Neurocomputing, vol.73, no.7–9, pp.1352–1361, 2010.

[22] C. Wang, J. Zhao, X. He, C. Chen, and J. Bu, "Image retrieval using nonlinear manifold embedding," Neurocomputing, vol.72, no.16–18, pp.3922–3929, 2009.

[23] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and Schölkopf, "Learning with local and global consistency," in Advances in Neural Information Processing Systems, pp.595–602, MIT Press, 2004.

[24] X.S. Zhou, A. Garg, and T.S. Huang, "Nonlinear variants of biased discriminants for interactive image retrieval," IEE Proc.-Vision, Image and Signal Processing, pp.927–936, 2005.

[25] J.F. Bonnans and A. Shapiro, "Optimization problems with perturbations: A guided tour," SIAM review, vol.40, no.2, pp.228–264, 1998.

[26] R.A. Horn and C.R. Johnson, Matrix Analysis, Cambridge University Press, 1990.

[27] A. Vadivel, A.K. Majumdar, and S. Sural, "Image retrieval using wavelet based texture features," International Conference on Communications, Devices and Intelligent Systems, pp.608–611, Kolkata, India, 2004.

Xu Yang received the B.E. and M.E. degrees in Information and Control Engineering from Liaoning Shihua University, Fushun, China, in 2001 and 2004, respectively. He is currently a Ph.D. student in the Institute of Image Processing and Pattern Recognition at Shanghai Jiao Tong University Shanghai, China. His research interests include machine learning and medical image analysis.

Huilin Xiong received the B.Sc. and M.Sc. degrees in Mathematics from Wuhan University, Wuhan, China, in 1985 and 1988, respectively. He received his Ph.D. degree in Pattern Recognition and Intelligent Control from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1999. He joined Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2007, and currently, he is a professor in the Department of Automation of SJTU. His research interests include kernel-based nonlinear pattern recognition, machine learning, and bioinfomatics.

Xin Yang received the M.Sc. degree in control engineering from Northwestern Polytechnic University, Xian, China, in 1982 and the Ph.D. degree of Applied Science degree in electronic engineering from the Free University of Brussels (ETRO/VUB) in 1995. Since 1997, he has been with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His current research activities are in the area of medical image analysis and partial differential equations in image processing. Dr. Yang is the author of more than 70 papers in refereed journals and conference proceedings.