*Journal of Information Hiding and Multimedia Signal Processing* ©2013 ISSN 2073-4212
**Ubiquitous International** Volume **4**, Number **4**, October **2013**

# Optimizing Kernel Function with Applications to Kernel Principal Analysis and Locality Preserving Projection for Feature Extraction

Jiaqing Qiao, Hongtao Yin

Department of Automatic Test and Control
Harbin Institute of Technology, No. 92 Xidazhi Street, Harbin 150001, China

Received March, 2013; revised June, 2013

Abstract. *Kernel learning is a popular research topic in pattern recognition and machine learning. Kernel selection is a crucial problem endured by kernel learning method in the practical applications. Many methods of finding the optimal parameters have been presented, but this kind of methods have no ability of changing the kernel structure, accordingly without changing the data distribution in kernel mapping space. In this paper, we present a uniform framework of kernel optimization based on data-dependent kernel from theory to applications to kernel principal analysis and locality preserving projection for feature extraction. Some experiments are implemented to evaluate the performance and feasibility of this framework.*
**Feature extraction, machine learning, kernel principal analysis, locality preserving projection**

1. **Introduction.** On current kernel learning methods, the performance of many linear learning methods is improved because the data distribution in the nonlinear feature space is easy to classification owing to kernel mapping. The geometrical structure of the data in the kernel mapping space, which is totally determined by the kernel function, has significant impact on the performance of these kernel learning methods. The discriminative ability of the data in the feature space could be even worse if an inappropriate kernel is used. Moreover, researchers optimized the parameters of kernel function to improve KDA [1] [2], [3], but these methods only choosing the optimal parameter of kernel from a set of discrete values which are created in advance. The geometry structure of data distribution in the kernel space is not be changed only through the changing the parameters of kernel. Xiong proposed a data-depend kernel for kernel optimization [4], and Amari presented support vector machine classifier through modifying the kernel function [6]. In the previous works [2][5], authors present data-dependent kernel based KDA algorithm for face recognition application. In recent research, learning based methods are used in many areas, such as object recognition [9],[10]. Face detection [11],[12]. Image analysis [13]. Some algorithms using the kernel trick are developed in recent years, such as kernel principal component analysis (KPCA), kernel discriminant analysis (KDA) and support vector machine (SVM). KPCA was originally developed by Scholkopf et al. in 1998, while KDA was firstly proposed by Mika et al. in 1999. KDA has been applied in many real-world applications owing to its excellent performance on feature extraction. Researchers have developed a series of KDA algorithms (Juwei Lu[15], Baudat and Anouar [16], Liang and

Shi [17],[18],[19], Yang [36],[21], J. Lu [20], Zheng [22], Huang [23], Wang [24] and Chen [25], Yixiong Liang [26], Yu-jie Zheng [27], Dacheng Tao[28], Yong Xu [29], Kamel Saadi [30], Dit-Yan Yeung [31], LinLin Shen [32], Bo Ma [33], Xiao-Hong Wu [34], Qingshan Liu[35]).

As above discussion, kernel learning is an important research topic in the machine learning area, and some theory and applications fruits are achieved and widely applied in pattern recognition, data mining, computer vision, image and signal processing areas. The nonlinear problems are solved with kernel function, and system performances such as recognition accuracy, prediction accuracy are largely increased. However, kernel learning method still endures a key problem, i.e., kernel function and its parameter selection. Kernel function and its parameters have the direct influence on the data distribution in the nonlinear feature space, and the inappropriate selection will influence the performance of kernel learning. In this paper, we focus on two schemes: one is kernel optimization algorithm and procedure, and the other is the framework of kernel learning algorithms. To verify the effectiveness of the kernel optimization scheme proposed, the proposed kernel optimization method is applied into popular kernel learning methods including kernel principal component analysis, kernel discriminant analysis and kernel locality preserving projection.

In the rest paper, firstly we have analyzed the trends in kernel learning algorithms and presented the kernel learning methods including Sparse KPCA and KCLPP, and secondly we present the theoretical derivation and algorithm procedure of kernel self-optimization learning, and thirdly we have the applications on Sparse KPCA and KCLPP. Finally the comprehensive experimental comparison and analysis are implemented to testify the performance of kernel self-optimization on simulated data, UCI dataset, and ORL and YALE databases.

## 2. **Theory.**

### 2.1. **Framework.**

We still employ the data-dependent kernel to improve the recognition accuracy, but owing to consider the computing problem the constraint optimization equation of solving the adaptive parameter of data-dependent kernel function must be different from the previous work [4] and changed through considering the computing problem. Data-dependent kernel is defined as

$$k(x,y) = f(x)f(y)k_0(x,y) \tag{1}$$

where $k_0(x,y)$ is the basic kernel. Polynomial kernel and Gaussian kernel can be the basic kernel. $f(x)$ is the positive value of $x$, the data-dependent kernel with the different $f(x)$ have the different performance. $f(x)$ is defined with $f(x) = \sum_{i\in SV} a_i e^{-\delta\|x-\tilde{x}_i\|^2}$, where $\tilde{x}_i$ is the $i$th support vector, $SV$ is the support vector, $a_i$ represents the contribution of $\tilde{x}_i$, $\delta$ is the free value.

SO

$$f(x) = a_0 + \sum_{n=1}^{N_{XV}} a_n e(x,\tilde{x}_n) \tag{2}$$

where $\delta$ is the free parameter, $\tilde{x}_i$ is the Expansion Vectors ($XVs$), $N_{XV}$ is the number of expansion vectors, and $a_n (n = 0, 1, 2, \cdots, N_{XVs})$ are the corresponding Expansion Coefficients. According to the extended definition of data-dependent kernel function, supposed the free parameter $\delta$ and expansion vector $\tilde{x}_n (n = 0, 1, 2, ..., N_{XV_s})$, the geometry structure of the data in the nonlinear mapping projection space is changeable with the changing of

the expansion coefficient $\alpha_n (n = 0, 1, 2, ..., N_{XV_s})$. So we can adjust the geometry structure of data in the nonlinear mapping space through changing the expansion coefficient. In order to optimize the kernel function through finding the optimal expansion coefficient, consider the computation problem, we optimize the kernel function in the Empirical Feature Space to solve the objective function through maximizing discriminantion.

2.2. **Optimization.** In this section, we present a novel kernel optimization objection function based two criterions, Fisher criterion and maximum margin criterion.

**(1) Fisher criterion based objective function**

Under the different expansion coefficient vector $\alpha$, the geometry structure of data in the empirical space causes the discriminative ability of samples. There are many methods of solving the above optimization equation. Supposed that $\alpha$ is an unit vector, i.e., $\alpha^T \alpha = 1$. ,the constrained equation is created to solve the optimal $\alpha$ as follows.

$$\max \quad J_{Fisher}(\alpha)$$
$$subject \quad to \quad \alpha^T \alpha - 1 \ = 0 \tag{3}$$

where $J_{Fisher}(\alpha)$ is a function with its variable $\alpha$ as defined as follows

$$J_{Fisher}\,(\alpha) = (\alpha^T E^T B_0 E \alpha)/(\alpha^T E^T W_0 E \alpha) \tag{4}$$

where $E^T B_0 E$ and $E^T W_0 E$ are constant matrix.

The objective function, Fisher criterion is to measure the class discriminative ability of the samples in the empirical feature space. $J_{Fisher} = tr(S_B^\Phi)/tr(S_W^\Phi)$ measures the discriminative ability of samples in the empirical feature space, where $J_{Fisher}$ measure the linear discriminative ability, $S_B^\Phi$ is the between class scatter matrix, $S_W^\Phi$ is inter class scatter matrix , and $tr$ denotes the trace. Let $k$ is the kernel matrix with its element $k_{ij}\,(i, j = 1, 2, ..., n)$ is calculated with $x_i$ and $x_j$. The matrix $K_{pq}$, $p, q = 1, 2, ..., L$ is the $n_p \times n_q$ matrix with p and q class. Then in the empirical feature space, we can obtain $tr(S_B^\Phi) = 1_n^T B 1_n$ and $tr(S_W^\Phi) = 1_n^T W 1_n$, where $B = diag(\frac{1}{n_1}K_{11}, \frac{1}{n_2}K_{22}, ..., \frac{1}{n_L}K_{LL}) - \frac{1}{n}K$. The class discriminative ability is defined as

$$J_{Fisher} = (1_n^T B 1_n)/(1_n^T W 1_n) \tag{5}$$

According to the definition of the data-dependent kernel , let $D = diag(f(x_1),$ $f(x_2), ..., f(x_n))$, the relation between the data-dependent kernel matrix $K$ and the basic kernel matrix $K_0$ calculated with basic kernel function $k_0(x, y)$ is defined as

$$K = DK_0D \tag{6}$$

Then $J_{Fisher} = (1_n^T D B_0 D 1_n)/(1_n^T D W_0 D 1_n)$, where $l_n$ is $n$ dimensional unit vector, according to the definition of data-dependent kernel, then $D1_n = E\alpha$, where $\alpha = [a_0, a_1, a_2, ..., a_{N_{XVs}}]^T$. On the solution of the objective function, we solve the objective function as follows. The following method is a classic method. Let $J_1(\alpha) = \alpha^T E^T B_0 E \alpha$ and $J_2(\alpha) = \alpha^T E^T W_0 E \alpha$, then

$$\begin{cases} \frac{\partial J_1(\alpha)}{\alpha} = 2E^T B_0 E \alpha \\ \frac{\partial J_2(\alpha)}{\alpha} = 2E^T W_0 E \alpha \end{cases} \tag{7}$$

Then $\frac{\partial J_{Fisher}(\alpha)}{\partial \alpha} = \frac{2}{J_2^2}(J_2 E^T B_0 E - J_1 E^T W_0 E)\alpha$, let $\frac{\partial J_{Fisher}(\alpha)}{\partial \alpha} = 0$, then

$$J_1 E^T W_0 E\alpha = J_2 E^T B_0 E\alpha \tag{8}$$

If $\left(E^T W_0 E\right)^{-1}$ exists, then

$$J_{Fisher}\alpha = (E^T W_0 E)^{-1}(E^T B_0 E)\alpha \tag{9}$$

$J_{Fisher}$ is equal to the eigenvalue of $(E^T W_0 E)^{-1}(E^T B_0 E)$, and the corresponding eigenvector is equal to expansion coefficients vector $\alpha$. In many applications, the matrix $(E^T W_0 E)^{-1}(E^T B_0 E)$ is not symmetrical or $E^T W E$ is singular. So the iteration method is used to solve the optimal $\alpha$, that is

$$\alpha^{(n+1)} = \alpha^{(n)} + \varepsilon(\frac{1}{J_2}E^T B_0 E - \frac{J_{Fisher}}{J_2}E^T W_0 E)\alpha^{(n)} \tag{10}$$

$\varepsilon$ is the learning rate, $\varepsilon\,(n) = \varepsilon_0(1 - \frac{n}{N})$, where $\varepsilon_0$ is the initialized learning rate, $n$ and $N$ is the current iteration number and the total iteration number in advance respectively.

The initialized learning rate $\varepsilon_0$ and the total iteration number $N$ is set in advance for the solution of the expansion coefficient. The initial learning rate $\varepsilon_0$ influences the convergence speed of the algorithm, and the total iteration number $N$ determines the time of solution. Only when the parameter $\varepsilon_0$ and $N$ are chosen appropriately we choose the optimal expansion coefficient vector. So the solution of expansion coefficient is not unique, which is determined by the selection of learning parameter. The iteration algorithm costs much time.

**(2) Maximum margin criterion (MMC) based objective function**

Based on maximum margin criterion, let $\alpha^T\alpha = 1$, the objective function of kernel optimization is defined as

$$\max \quad \alpha^T(2\tilde{S}_B - \tilde{S}_T)\alpha$$

$$subject \quad to \quad \alpha^T\alpha - 1 \ = 0 \tag{11}$$

It is easy to know that the optimal expansion coefficient vector $\alpha^*$ is equal to the eigenvector of $2\tilde{S}_B - \tilde{S}_T$ corresponding to the maximal eigenvalue. Where $\begin{cases} \tilde{S}_B = X_B X_B^T \\ \tilde{S}_T = X_T X_T^T \end{cases}$ and $\begin{cases} X_T = (Y_0 - \frac{1}{m}Y_0 1_m^T 1_m)E \\ X_B = Y_0 M^T E \end{cases}$, where $M = M_1 - M_2$, and $M_1, M_2$ are defined as

$$M_1 = \begin{bmatrix} \left[\frac{1}{\sqrt{m_1}}\right]_{m_1\times m_1} & 0_{m_1\times m_2} & \cdots & 0_{m_1\times m_c} \\ 0_{m_2\times m_1} & \left[\frac{1}{\sqrt{m_2}}\right]_{m_2\times m_2} & \cdots & 0_{m_2\times m_c} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{m_c\times m_1} & 0_{m_c\times m_2} & \cdots & \left[\frac{1}{\sqrt{m_c}}\right]_{m_c\times m_c} \end{bmatrix} \text{ and}$$

$$M_2 = \begin{bmatrix} \frac{\sum_j^c \sqrt{m_j}}{m} & 0 & \cdots & 0 \\ 0 & \frac{\sum_j^c \sqrt{m_j}}{m} & & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sum_j^c \sqrt{m_j}}{m} \end{bmatrix} \cdot Y_0 = K_0 P_0 \Lambda_0^{-1/2}.$$

$Y_0 = K_0 P_0 \Lambda_0^{-1/2}$, $K_0 = P_0 \Lambda_0^T P_0^T$, and $K_0$ is the basic kernel matrix. The objective function is based on maximum margin criterion, which is defined as

$$Dis = \frac{1}{2n} \sum_{i=1}^{L} \sum_{j=1}^{L} n_i n_j d\left(c_i, c_j\right) \tag{12}$$

where $d(c_i, c_j) = d(m_i^\Phi, m_j^\Phi) - S(c_i) - S(c_j)$ denotes the margin between class $i$ and class $j$, and $d(m_i^\Phi, m_j^\Phi)$ denotes the distance between the centers of two classes of samples, and $S(c_i)$ denotes the scatter matrix $c_i (i = 1, 2, ..., L)$ is defined as where $tr(S_i^\Phi) = \frac{1}{n_i} \sum_{p=1}^{n_i} (\Phi(x_i^p) - m_i^\Phi)^T (\Phi(x_i^p) - m_i^\Phi)$, $S_i^\Phi$ is the scatter matrix of $i$th class. Then $Dis = tr(S_B^\Phi) - tr(S_W^\Phi)$. It is easy to obtain $Dis = tr(2S_B^\Phi - S_T^\Phi)$. In the empirical feature space, the sample set $Y = KP\Lambda^{-1/2}$, wher $K$ is the data-dependent kernel. $P$ and $\Lambda$ satisfies $K = P\Lambda^T P^T$, then

$$S_T = (Y - \frac{1}{m} Y 1_m^T 1_m)(Y - \frac{1}{m} Y 1_m^T 1_m)^T \tag{13}$$

where $1_m = [1, 1, ..., 1]_{1 \times m}$, with $D1_n = E\alpha$, then

$$trace(S_T) = \alpha^T ((Y_0 - \frac{1}{m} Y_0 1_m^T 1_m) E)^T ((Y_0 - \frac{1}{m} Y_0 1_m^T 1_m) E) \alpha \tag{14}$$

Let $X_T = (Y_0 - \frac{1}{m} Y_0 1_m^T 1_m) E$, then

$$trace(S_{_T}) = \alpha^T (X_T)^T X_T \alpha \tag{15}$$

where $Y_0 = K_0 P_0 \Lambda_0^{-1/2}$, $K_0 = P_0 \Lambda_0^T P_0^T$, and $K_0$ is the basic kernel matrix. Similarly, it easy to obtain $S_B = (\sqrt{m_1}(u_1 - u), ..., \sqrt{m_c}(u_c - u))(\sqrt{m_1}(u_1 - u), ..., \sqrt{m_c}(u_c - u))^T$, where $1_c = [1, 1, ..., 1]_{1 \times c}$. Supposed that $M = M_1 - M_2$, and

$$M_1 = \begin{bmatrix} \left[\frac{1}{\sqrt{m_1}}\right]_{m_1 \times m_1} & 0_{m_1 \times m_2} & \cdots & 0_{m_1 \times m_c} \\ 0_{m_2 \times m_1} & \left[\frac{1}{\sqrt{m_2}}\right]_{m_2 \times m_2} & \cdots & 0_{m_2 \times m_c} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{m_c \times m_1} & 0_{m_c \times m_2} & \cdots & \left[\frac{1}{\sqrt{m_c}}\right]_{m_c \times m_c} \end{bmatrix} \text{ and}$$

$$M_2 = \begin{bmatrix} \frac{\sum_j^c \sqrt{m_j}}{m} & 0 & \cdots & 0 \\ 0 & \frac{\sum_j^c \sqrt{m_j}}{m} & & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sum_j^c \sqrt{m_j}}{m} \end{bmatrix}.$$

So $trace(S_B) = \alpha^T (Y_0 M^T E)^T (Y_0 M^T E) \alpha$, with $X_B = Y_0 M^T E$, then

$$\begin{cases} trace(S_B) = \alpha^T X_B^T X_B \alpha \\ trace(S_{_T}) = \alpha^T (X_T)^T X_T \alpha \end{cases} \tag{16}$$

Let $\tilde{S}_B = X_B X_B^T$ and $\tilde{S}_T = X_T X_T^T$, then

$$\widetilde{Dis}(\alpha) = trace(\alpha^T(2\tilde{S}_B - \tilde{S}_T)\alpha) \tag{17}$$

So, maximize $\widetilde{Dis}(\alpha)$ is equal to obtain the objective function in (11). Solving the objective function in (29) through calculating the eigenvector and eigenvalue of matrix $2\tilde{S}_B - \tilde{S}_T$ as $P^T\tilde{S}_BP = \Lambda$ and $P^T\tilde{S}_TP = I$, where $P = \phi\theta^{-1/2}\psi$, $\theta$ and $\phi$ are the eigenvalue and eigenvector of $\tilde{S}_T$ respectively. $\psi$ is the eigenvalue matrix of $\theta^{-1/2}\phi^T\tilde{S}_B\phi\theta^{-1/2}$. So the column vector of $P$ is the eigenvalue matrix of $2\tilde{S}_B - \tilde{S}_T$ corresponding to the eigenvalue $2\Lambda - I$.

**(3) Discussion**

Differences between two kernel optimization methods of Fisher criterion and maximum margin criterion are shown as follows. Fisher criterion method use iteration method to solve the solution, while maximum margin criterion method is to find the optimal solution with the eigenvalue equation. Fisher criterion method cost much more time than maximum margin criterion method. Moreover, Fisher criterion method needs to choose the relative parameters in advance, while maximum margin criterion method needs not to choose the parameters in advance.

## 3. Applications.

3.1. **Application to Sparse KPCA.** In this section, we apply kernel optimization method to Sparse KPCA [7]. Sparse KPCA is formulated with the viewpoint of least squares support vector machine. Sparse KPCA endures two problems, one is that all training samples need to be stored for the computing the kernel matrix during kernel learning, and the second is that the kernel and its parameter have the heavy influence on performance of kernel learning. We apply the kernel function $k'(x, y)$ into Sparse KPCA as follows.

$$y = B^TV_{zx} \tag{18}$$

where $g(z_i, x) = k'(z_i, x) - \frac{1}{N}\sum_{q=1}^{N} k'(z_i, x_q)$, $V_{zx} = \left[\begin{array}{cccc} g(z_1, x) & g(z_2, x) & \ldots & g(z_{N_z}, x) \end{array}\right]^T$. So

$$y = \sum_{i=1}^{N_z} \beta_i^z \left[\phi(z_i)^T\left(\phi(x) - u^\phi\right)\right] \tag{19}$$

Let $\beta_z = \left[\begin{array}{cccc} \beta_1^z & \beta_2^z & \cdots & \beta_{N_z}^z \end{array}\right]^T$. For we choose $m$ eigenvector $\alpha$ corresponding to $m$ largest eigenvalue. Let $P = \left[\begin{array}{cccc} \left(\beta_z{}^T\right)_1 & \left(\beta_z{}^T\right)_2 & \cdots & \left(\beta_z{}^T\right)_m \end{array}\right]^T$, the feature can be obtained as follows.

$$z = PK_{zx} \tag{20}$$

As above discussion from the theoretical viewpoints, kernel-optimized Sparse KPCA chooses adaptively a few of samples from the training sample set but a little influence on recognition performance, which saves much space of storing training samples on computing the kernel matrix with the lower time consuming. So in the practical applications, kernel-optimized Sparse KPCA solves the limitation from KPCA owing to its high store space and time consuming its ability on feature extraction. So from the theory viewpoint, this application of kernel-optimized learning is adaptive to the applications with the demand of the strict computation efficiency but not strict on recognition.

3.2. **Application to KLPP.** Kernel CLPP method is presented in our previous work[10]. The kernel LPP are shown as follows. With the constraint $\left(Z^{\Phi}\right)^{T} D^{\Phi} Z^{\Phi} = 1$, i.e., $\alpha^{T} K D^{\Phi} K \alpha = 1$, the objective function is defined follows.

$$\min_{\alpha} \alpha^{T} K L^{\Phi} K \alpha$$

$$Subject to \alpha^{T} K D^{\Phi} K \alpha = 1 \tag{21}$$

where $L^{\Phi} = D^{\Phi} - S^{\Phi}$. We apply the kernel optimization method into Kernel CLPP. With kernel optimization method, we obtain the optimal $\beta^{*}$ of optimized data-dependent kernel. With this optimized kernel, the objective function of KLPP is defined as

$$\min_{\alpha} \alpha^{T} K^{(\beta^{*})} L^{\Phi} K^{(\beta^{*})} \alpha$$

$$Subject to \alpha^{T} K^{(\beta^{*})} D^{\Phi} K^{(\beta^{*})} \alpha = 1 \tag{22}$$

where the optimal projection $\alpha$ is the main projection vector to construct the projection matrix. Supposed that $r_1, r_2, ..., r_m$ are $K$'s orthonormal eigenvector corresponding to $m$ largest nonzero eigenvalue $\lambda_1, \lambda_2, ..., \lambda_m$. i.e., $K = P \Lambda P^{T}$ with QR decomposition, where $P = [r_1, r_2, ..., r_m]$ and $\Lambda = diag\,(\lambda_1, \lambda_2, ..., \lambda_m)$. The solution of the above constrained optimization problem is equal to the eigenvector corresponding to the largest eigenvalue.

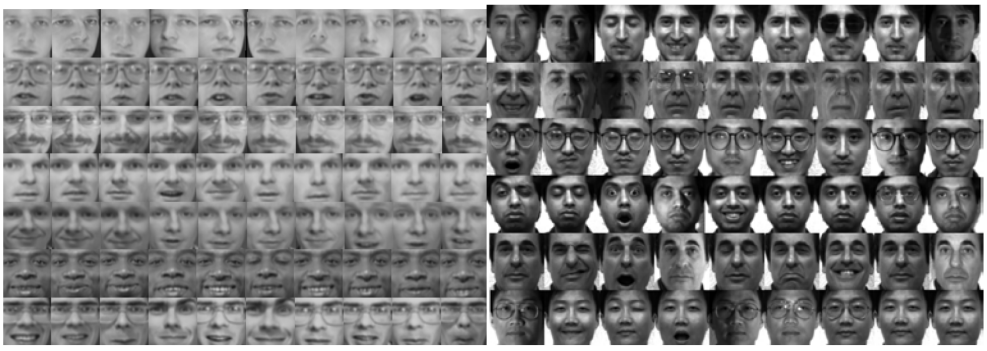
4. **Results.** On UCI dataset, we evaluate the feasibility and performance of kernel optimization using Sparse KPCA. In the experiments, we use the sparse analysis to determine some key training samples as the final sample for kernel learning. In our previous work [11], we have concluded the good recognition performance be achieved only using the less size of key training samples. For a higher performance, kernel optimization is introduced to the recognition performance, which is shown in Table 1. With same size of training samples, the kernel optimization based SKPCA performs better than KPCA. It is meaningful to achieve a higher performance but little size of training samples, which saves the saving space of training samples and time consuming for the applications with a large size of training samples. SKPCA saves much space of storing training samples on computing the kernel matrix with the lower time consuming, but achieves the similar recognition accuracy compared with KPCA. Kernel optimization based SKPCA algorithm achieves the higher recognition accuracy than SKPCA owing to its kernel optimization combined with SKPCA, which is adaptive to the applications with the demand of the strict computation efficiency but not strict on recognition.

The second set of experiments are implemented on real databases, including ORL and Yale databases. ORL face database is composed of 400 grayscale images with 10 images for each of 40 individuals with the variations in across pose and facial expression. Yale face database contains 165 grayscale images of 15 individuals, and these images are taken under different lighting condition (left-light, center-light, and right-light), and different facial expression (normal, happy, sad, sleepy, surprised, and wink), and with/without glasses. Since in the practical application, the face detection is the first step for recognition. We sized the image to pixels for ORL and for Yale to simulate the face detection in the real system. Secondly, some remarks on the experiment setting should be emphasized after the description of two databases in our experiments. For two databases, we randomly choose 5 images as the training images and the rest as test ones and use the average result of 10 times of experiments as the final recognition accuracy. to evaluate the performance.

On ORL and Yale database, we implement kernel-optimized learning version of KDA and KCLPP compared with the traditional LPP [8] together with PCA, KDA, CLPP[8] and KCLPP [8]. We use Fisher classifier for classification and implement the algorithms

TABLE 1. Performance of kernel optimization on UCI database (%)

| Datasets | Training samples | Key samples | SKPCA | Kernel optimization-SKPCA |
|---|---|---|---|---|
| Banana | 400 | 120 | 14.2±0.1 | 13.9±0.2 |
| Image | 1300 | 180 | 5.4±0.3 | 5.1±0.2 |
| F.Solar | 666 | 50 | 34.2±2.3 | 32.8±2.1 |
| Splice | 1000 | 280 | 9.4±0.9 | 9.0±0.7 |
| Thyroid | 140 | 30 | 2.2±1.3 | 1.8±1.0 |
| Titanic | 150 | 30 | 23.2±0.5 | 22.4±0.4 |



(a) ORL (b) Yale

FIGURE 1. Example face images of face databases used in our experiments.

form many times and the averaged recognition rate is considered as the recognition accuracy. As shown in Table 2, the averaged recognition rate of LPP, CLPP, KCLPP and the proposed kernel-optimized KCLPP are 93.80%, 94.80%, 96.50% and 98.00% respectively. Kernel-optimized KCLPP achieves the highest recognition accuracy through using kernel self-optimization because the data structure has changed according to the input data. On the Yale database, the performances are shown in Table 3.

As results shown in Table 2 and Table 3, kernel-optimized learning methods can obtain the higher recognition accuracy compared with their traditional kernel learning methods. Choosing of kernel function and its parameter is a key impact factor on recognition performance on kernel learning. The adaptively parameter choosing of data-dependent kernel function can improve the recognition performance of kernel learning under the same condition of the same number of training and test samples. Under the condition of the limited training samples stored in the databases, kernel-optimized learning methods are applied to increase the recognition accuracy with the same training samples compared with the traditional kernel learning. Besides the excellent recognition performance, the efficiency of kernel-optimized learning algorithms is still one problem worth to discuss.

TABLE 2. Recognition Accuracy on ORL Sub-Databases (%)

| Datasets | PCA[1] | KDA[2] | Optimized KDA | LPP[8] | CLPP[8] | KCLPP[8] | Optimized KCLPP |
|---|---|---|---|---|---|---|---|
| SD1 | 92.00 | 93.50 | 94.50 | 95.00 | 96.00 | 96.50 | 98.50 |
| SD 2 | 92.00 | 93.50 | 94.50 | 93.50 | 94.00 | 95.50 | 97.50 |
| SD 3 | 93.00 | 94.00 | 95.50 | 95.50 | 97.00 | 98.50 | 99.50 |
| SD4 | 92.00 | 93.50 | 94.50 | 93.50 | 94.50 | 96.00 | 97.50 |
| SD 5 | 90.50 | 91.00 | 92.50 | 91.50 | 92.50 | 96.00 | 97.00 |
| Averaged | 91.90 | 93.10 | 94.30 | 93.80 | 94.80 | 96.50 | 98.00 |

TABLE 3. Recognition Accuracy on YALE Sub-Databases (%)

| | PCA[1] | KDA[2] | Optimized KDA | LPP[8] | CLPP[8] | KCLPP[8] |
|---|---|---|---|---|---|---|
| YALE_SD1 | 84.33 | 94.44 | 86.33 | 90.00 | 94.44 | 95.67 |
| YALE_SD 2 | 86.77 | 92.22 | 90.67 | 91.11 | 92.22 | 93.33 |
| YALE_SD 3 | 85.33 | 93.33 | 88.56 | 86.67 | 93.33 | 94.44 |
| YALE_SD4 | 85.67 | 93.33 | 88.89 | 90.00 | 93.33 | 92.33 |
| YALE_SD 5 | 88.67 | 96.67 | 95.56 | 93.33 | 96.67 | 97.44 |
| Averaged | 86.15 | 94.00 | 90.00 | 90.22 | 93.99 | 94.62 |

Kernel-optimized learning methods cost much computing time on calculating the projection matrix. But in many applications, in order to achieve a higher computing recognition performance but less consideration of time consuming. In many practical applications, such as face recognition, it does not cause significantly increasing of the response time. So, kernel optimization is an effective way of improving recognition performance of kernel learning in the practical applications.

5. **Conclusions.** We present one kernel optimization method for kernel-based learning to solve the kernel function and its parameter selection problem, and this method has the same important practical meaning for the improving of kernel-based system. This paper presents a kernel optimization method with data dependent kernel based on Fisher criterion and maximum margin criterion, and then applies this method into other kernel learning methods.

**Acknowledgment.** This work is supported by "the Fundamental Research Funds for the Central Universities" (Grant No. HIT. NSRIF. 2013022).

## REFERENCES

[1] J. Huang, P. C. Yuen, W. S. Chen, and J. H. Lai, Kernel subspace LDA with optimized kernel parameters on face recognition. *Proc. of the 6th IEEE international conference on Automatic face and gesture recognition*, pp. 327-332, 2004.

[2] J. S. Pan, J. B. Li, and Z. M. Lu. Adaptive quasiconformal kernel discriminant analysis, *Journal of Neurocomputing*, vol. 71. no. 13-15, pp. 2754-2760, 2006.

[3] W. S. Chen, P. C. Yuen, J. Huang, and D. Q. Dai, Kernel machine-based one-parameter regularized fisher discriminant method for face recognition, *IEEE Trans. Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 35, no. 4, pp. 659-669, 2005.

[4] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, Optimizing the kernel in the empirical feature space, *IEEE Trans. Neural Networks*, vol. 16, no. 2, pp. 460-474, 2005.

[5] J. B. Li, J. S. Pan, and Z. M. Lu, Kernel optimization-based discriminant analysis for face recognition, *Journal of Neural Computing and Applications*, vol. 18, no. 6, pp. 603-612, 2009.

[6] S. Amari, and S. Wu, Improving support vector machine classifiers by modifying kernel functions, *Journal of Neural Networks*, vol. 12, no. 6, pp. 783-789, 1999.

[7] J. B. Li, L. J. Yu, and S. H. Sun, Refined kernel principal component analysis based feature extraction, *Chinese Journal of Electronics*, vol. 20, no. 3, pp. 467-470, 2011.

[8] J. B. Li, J. S. Pan, and S. C. Chu, Kernel class-wise locality preserving projection, *Journal of Information Sciences*, vol. 178, no. 7, pp. 1825-1835, 2008.

[9] S. Krinidis, and I. Pitas, Statistical analysis of human facial expressions, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 3, pp. 241-260, 2010.

[10] C. F. Lee, and W. T. Chang, Recovery of color images by composed associative mining and edge detection, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 4, pp. 310-324, 2010.

[11] H. G. Kaganami, S. K. Ali, and Z. Beiji, Optimal approach for texture analysis and classification based on wavelet transform and neural network, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 1, pp. 33-40, 2011.

[12] W.C. Hu, C.Y. Yang, D.Y. Huang, and C.H. Huang, Feature-based face detection against skin-color like backgrounds with varying illumination, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 2, pp. 123-132, 2011.

[13] D. Y. Huang, C. J. Lin, and W. C. Hu, Learning-based face detection by adaptive switching of skin color models and adaBoost under varying illumination, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 3, pp. 204-216, 2011.

[14] P. Puranik, P. Bajaj, A. Abraham, P. Palsodkar, and A. Deshmukh, Human perception-based color image segmentation using comprehensive learning particle swarm optimization, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 3, pp. 227-235, 2011.

[15] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 117-126, 2003.

[16] G. Baudat, and F. Anouar, Generalized discriminant analysis using a kernel approach, *Journal of Neural Computation*, vol. 12, no. 10, pp. 2385-2404, 2000.

[17] Z. Liang, and P. Shi, Uncorrelated discriminant vectors using a kernel method, *Journal of Pattern Recognition*, vol. 38, no. 2, pp. 307-310, 2005.

[18] Z. Liang, and P. Shi, Efficient algorithm for kernel discriminant anlaysis, *Journal of Electronics Letters*, vol. 40, no. 25, pp. 1579-1581, 2004.

[19] Z. Liang, and P. Shi, An efficient and effective method to solve kernel Fisher discriminant analysis, *Journal of Neurocomputing*, vol. 61, pp. 485-493, 2004.

[20] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, Face Recognition Using Kernel Direct Discriminant Analysis Algorithms, *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 117-126, 2003.

[21] M. H. Yang, Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods, *Proc. of The Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 215-220, 2002.

[22] W. Zheng, C. Zou, and L. Zhao, Weighted maximum margin discriminant analysis with kernels, *Journal of Neurocomputing*, vol. 67, pp. 357-362, 2005.

[23] J. Huang, P. C. Yuen, W. S. Chen, and J. H. Lai. Kernel subspace LDA with optimized kernel parameters on face recognition, *Proc. of The 6th IEEE international conference on Automatic face and gesture recognition*, pp. 327-332, 2004.

[24] L. Wang, K. L. Chan, and P. Xue, A criterion for optimizing kernel parameters in KBDA for image retrieval, *IEEE Trans. Systems, Man and Cybernetics-Part B: Cybernetics*, vol. 35, no. 3, pp. 556-562, 2005.

[25] W. S. Chen, P. C. Yuen, J. Huang, and D. Q. Dai, Kernel machine-based one-parameter regularized fisher discriminant method for face recognition, *IEEE Trans. Systems, Man and Cybernetics-Part B: Cybernetics*, vol. 35, no. 4, pp. 658-669, 2005.

[26] Y. Liang, C. Li, W. Gong, and Y. Pan, Uncorrelated linear discriminant analysis based on weighted pairwise fisher criterion, *Journal of Pattern Recognition*, vol. 40, no. 12, pp. 3606-3615, 2007.

[27] Y. J. Zheng, J. Yang, J. Y. Yang, and X. J. Wu, A reformative kernel Fisher discriminant algorithm and its application to face recognition, *Journal of Neurocomputing*, vol. 69, no. 13-15, pp. 1806-1810, 2006.

[28] 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. Multimedia*, vol. 8, no. 4, pp. 716-727, 2006.

[29] Y. Xu, D. Zhang, Z. Jin, M. Li, and J. Y. Yang, A fast kernel-based nonlinear discriminant analysis for multi-class problems, *Journal of Pattern Recognition*, vol. 39, no. 6, pp. 1026-1033, 2006.

[30] K. Saadi, N. L. C. Talbot, and G. C. Cawley, Optimally regularised kernel fisher discriminant classification, *Journal of Neural Networks*, vol. 20, no. 7, pp. 832-841, 2007.

[31] D. Y. Yeung, H. Chang, and G. Dai, Learning the kernel matrix by maximizing a KFD-based class separability criterion, *Journal of Pattern Recognition*, vol. 40, no. 7, pp. 2021-2028, 2007.

[32] L. L. Shen, L. Bai, and M. Fairhurst, Gabor wavelets and general discriminant analysis for face identification and verification, *Journal of Image and Vision Computing*, vol. 25, no. 5, pp. 553-563, 2007.

[33] B. Ma, H. Y. Qu, and H. S. Wong, Kernel clustering-based discriminant analysis, *Journal of Pattern Recognition*, vol. 40, no. 1, pp. 324-327, 2007.

[34] X. H. Wu, and J. J. Zhou, Fuzzy discriminant analysis with kernel methods, *Journal of Pattern Recognition*, vol. 39, no. 11, pp. 2236-2239, 2006.

[35] Q. Liu, H. Lu, and S. Ma, Improving kernel fisher discriminant analysis for face recognition, *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 42-49, 2004.