Kernel-based Weighted Discriminant Analysis with QR Decomposition and Its Application to Face Recognition

JIANQIANG GAO
Liaocheng University
School of Mathematical Sciences
Hunan Road NO.1, Liaocheng
P.R. CHINA
gaojianqiang82@126.com

LIYA FAN
Liaocheng University
School of Mathematical Sciences
Hunan Road NO.1, Liaocheng
P.R. CHINA
fanliya63@126.com

Abstract: Kernel discriminant analysis (KDA) is a widely used approach in feature extraction problems. However, for high-dimensional multi-class tasks, such as faces recognition, traditional KDA algorithms have a limitation that the Fisher criterion is non-optimal with respect to classification rate. Moreover, they suffer from the small sample size problem. This paper presents two variants of KDA called based on QR decomposition weighted kernel discriminant analysis (WKDA/QR), which can effectively deal with the above two problems, and based on singular value decomposition weighted kernel discriminant analysis (WKDA/SVD). Since the QR decomposition on a small size matrix is adopted, the superiority of the proposed method is its computational efficiency and can avoid the singularity problem. In addition, we compare WKDA/QR with WKDA/SVD under the parameters of weighted function and kernel function. Experimental results on face recognition show that the WKDA/QR and WKDA/SVD are more effective than KDA, and WKDA/QR is more effective and feasible than WKDA/SVD.

Key–Words: QR decomposition, Kernel discriminant analysis (KDA), Feature extraction, Face recognition, small sample size (SSS)

1 Introduction

Linear discriminant analysis (LDA), seeking optimal linear projections such that the Fisher criterion of the between-class scatter versus the within-class scatter is maximized, is one of the most well-known statistical technique for feature extraction and dimension reduction [1-4]. Recently, several extensions of LDA [5-8] have been developed concerning robustness issue. Although LDA is an effective method for feature extraction, it is still a linear technique in nature. Hence, it is not sufficient to deal with some features which have nonlinear relationships. To overcome the problem, the kernel trick is applied to effectively describe nonlinear relationships of input data. Recently, kernel-based learning methods have attracted much attention in the areas of pattern recognition and machine learning. Scholkopf et al. [9] applied the kernel trick to principal component analysis (KPCA), which can effectively compute principal components in the high-dimensional feature space. Mika et al. [10] proposed kernel discriminant analysis (KDA) for two-class cases. Baudat and Anouar [11] developed a generalized kernel discriminant analysis (GKDA) for multiclass problems. Because of its ability to extract discriminant nonlinear features, KDA has been used widely in many real-world applications such as document analysis, face recognition and image retrieval [12-16].

Yang et al. [16] further discussed kernel Fisher discriminant analysis and pointed out that kernel Fisher discriminant analysis is equivalent to kernel principal component analysis plus Fisher linear discriminant analysis. Therefore, for high-dimensional multi-class tasks such as faces recognition, the original KDA-based algorithms usually encounter three difficulties: the first is the singularity problem caused by the small sample size (SSS) problem [11-13], in which the number of training samples is far smaller than the dimension of the sample. Moreover, because KDA uses an implicit nonlinear mapping to project low-dimensional input patterns into a high-dimensional feature space, many large sample size problem in input space maybe changed into SSS problems in the feature space. The second is that the Fisher separability criterion is not directly related to classification rate, that is, the classes with larger distance to each other in feature space are more emphasized when the Fisher criterion is optimized, which leads that the resulting projection preserves the distance of already well-separated classes, causing a large overlap of neighboring classes [17-20]. The third is that these algorithms still face the computational difficulty of the eigen-decomposition of matrices in the high-

dimensional space. Because the three problems are common in many applications, it is necessary to develop new and more effective KDA algorithms to deal with them.

In fact, the same three problems are also appeared in LDA-based methods. fortunately, LDA has been well studied and many LDA extension algorithms have been proposed to deal with the problems. Lotlikar and Kothari [17] and Loog et al. [18] presented weighted versions of LDA for high-dimensional multi-class problem. Mika et al. [10,13] used a regularization technique that makes the inner product matrix be nonsingular by adding a scalar matrix. Baudat and Anouar [11] employed the QR decomposition technique to eliminate the singularity by removing the zero eigenvalues. Lu et al. [14] presented kernel direct discriminant analysis, which is a generalization of the direct-LDA [21]. Recently, Dai et al. [19,22], Zhou and Tang [23], and Zhou and Tang [24] presented kernel-weighted discriminant analysis by generalizing the fractional LDA [17]. The main methods in [23-24] are the simultaneous diagonalization technique for tackling the SSS problem.

Motivated by shortcomings of KIDA method in [23-24], this paper presents two new kernel-weighted discriminant analysis (WKDA) for feature extraction: WKDA with QR decomposition (WKDA/QR) and WKDA with SVD decomposition (WKDA/SVD). Experiments on face recognition task show that WKDA/QR and WKDA/SVD are more effective than KDA, and WKDA/QR is better than WKDA/SVD to nonlinear feature extraction.

The rest of this paper is organized as follows. The KDA method is briefly introduced and discussed in Section 2. The detailed descriptions of WKDA/QR and WKDA/SVD are presented in Section 3. In Section 4, the feature extraction performances of WKDA/QR and WKDA/SVD on face recognition task are reported by comparing them with typical KDA algorithm. Section 5 concludes the paper.

2 Review of KDA

Kernel discriminant analysis (KDA) is a kernel version of LDA to deal with the feature extraction and classification of nonlinear characteristics. The basic idea of KDA is to firstly project original patterns into a high-dimensional feature space F by an implicit nonlinear mapping $\phi : R^n \to F : x \to \Phi(x)$ and then to use LDA in feature space F.

Let us consider a set of m training samples $\{x_1, x_2, \cdots, x_m\}$ taking values in an n dimensional space. Let L be the number of classes and m_i the number of training samples in the i-th class, $i = 1, \cdots, L$. Obviously, $m = \sum_{i=1}^{L} m_i$. In general, the Fisher criterion [16,25] can be defined as

$$\max_v J(v) = \frac{v^T S_b^\phi v}{v^T S_t^\phi v}, \tag{1}$$

where $S_b^\phi = \frac{1}{m}\sum_{i=1}^{L} m_i(m_i^\phi - m_0^\phi)(m_i^\phi - m_0^\phi)^T$ and $S_t^\phi = \frac{1}{m}\sum_{i=1}^{m}(\phi(x_i) - m_0^\phi)(\phi(x_i) - m_0^\phi)^T$ are the between-class and total scatter matrixes defined in the feature space F, respectively, where m_i^ϕ is the mean vector of the mapped training samples in the i-th class and m_0^ϕ is the mean vector of all mapped training samples. The optimization problem (1) can be transformed into the following eigenvalue problem:

$$S_b^\phi v = \lambda S_t^\phi v. \tag{2}$$

Let $\Phi(X) = [\phi(x_1), \cdots, \phi(x_m)]$ and $k : R^n \times R^n \to R$ be a kernel function. The kernel matrix $K = (k_{ij}) \in R^{m\times m}$ corresponded to the kernel k can be defined by $k_{ij} = k(x_i, x_j) =< \phi(x_i), \phi(x_j) >$, where $\phi : R^n \to F$ is a feature map and F is a feature space of the kernel k. It is evident that $K = \Phi(X)^T\Phi(X)$. For any $j \in \{1, \cdots, m\}$, let $\widetilde{\phi}(x_j) = \phi(x_j) - \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)$ be the centered mapped data and $\widetilde{\Phi}(X) = [\widetilde{\phi}(x_1), \cdots, \widetilde{\phi}(x_m)] = \Phi(X)(I - 1_{m\times m}/m)$, where I is a $m \times m$ identity matrix and $1_{m\times m}$ is a $m \times m$ matrix of all ones. The inner product matrix $\widetilde{K}$ for the centered mapped data can be obtained by

$$\begin{aligned}\widetilde{K} &= \widetilde{\Phi}(X)^T\widetilde{\Phi}(X)\\ &= (I - 1_{m\times m}/m)^T K(I - 1_{m\times m}/m).\end{aligned} \tag{3}$$

According to the reproducing kernel theory [26], the eigenvector v lies in the span of $\{\widetilde{\phi}(x_1), \cdots, \widetilde{\phi}(x_m)\}$ and then there exist coefficients $a_i, i = 1, 2, \cdots, m$ such that

$$v = \sum_{i=1}^{m} a_i\widetilde{\phi}(x_i) = \widetilde{\Phi}(X)a, \tag{4}$$

where $a = (a_1, \cdots, a_m)^T$. Let $W = diag(s_1, \cdots, s_j, \cdots, s_L)$, where s_j is a $m_j \times m_j$ matrix whose elements are $1/m_j$. Substituting (4) into (1), we can obtain the following equation:

$$\max_a J(a) = \frac{a^T\widetilde{K}W\widetilde{K}a}{a^T\widetilde{K}\widetilde{K}a}. \tag{5}$$

In general, the vector a_1 corresponding to the maximal value of $J(a)$ is the optimal discriminant direction. However, in some cases, it is not enough to only use one optimal discriminant direction to feature extraction. Hence, it is often necessary to obtain t

$(t > 1)$ optimal discriminant directions. Assume that $a_1, \cdots, a_t$ are t optimal discriminant directions and $A = [a_1, a_2, \cdots, a_t]$. Then A should satisfy

$$A = \arg\max_A \mathrm{tr}(\frac{A^T S_b' A}{A^T S_t' A}), \tag{6}$$

where $S_b' = \widetilde{K} W \widetilde{K}$, $S_t' = \widetilde{K}\widetilde{K}$, and $tr(\cdot)$ denotes the trace of matrices. The optimization problem (6) can be transformed into the following generalized eigenvalue problems:

$$S_b' a = \lambda S_t' a. \tag{7}$$

The solution of the problem (7) can be obtained by solving the generalized eigenvalue problem. Suppose that $\lambda_1, \lambda_2, \cdots, \lambda_t$ are the t largest eigenvalues of the problem (7) sorted in descending order and $a_1, \cdots, a_t$ are the corresponding eigenvectors. We can obtain the KDA transform matrix by

$$\begin{aligned} V &= [v_1, \cdots, v_t] = \widetilde{\Phi}(X)[a_1, \cdots, a_t] \\ &= \widetilde{\Phi}(X)A. \end{aligned} \tag{8}$$

For any input vector x, its low-dimension feature representation y_x can be defined by

$$\begin{aligned} y_x &= V^T \widetilde{\phi}(x) \\ &= A^T \widetilde{\Phi}(X)^T \widetilde{\phi}(x) \\ &= A^T (\widetilde{k}(x_1, x), \widetilde{k}(x_2, x), \cdots, \widetilde{k}(x_m, x))^T. \end{aligned} \tag{9}$$

3 WKDA/QR and WKDA/SVD algorithms

In this section, two efficient and effective algorithms are proposed to solve kernel discriminant analysis, which are called WKDA/QR and WKDA/SVD for short. The main idea of WKDA/QR and WKDA/SVD is that original samples are projected firstly into a feature space of a kernel function by an implicit feature mapping and then use weighted LDA, where the between-class scatter matrix in the feature space is defined by pairwise weighted functions. In weighted LDA, the QR and SVD decomposition are employed to find low-dimensional nonlinear feature with significant discrimination power, respectively.

3.1. Pairwise weighted schemes

To obtain a modified criterion that it is more closely related to classification error, weighted schemes can be introduced into the traditional Fisher criterion to penalize the classes that are close in the feature space and then lead to potential misclassifications in the output space. However, we would like to keep the general form of Eq.(6), because in such form the following optimization can then be carried out by solving a generalized eigenvalue problem without having to resort to complex iterative optimization schemes. Therefore, pairwise weighted schemes, that pairs of classes with smaller distance, are introduced into the reconstruction of the between-class scatter matrix in the feature space.

Let $d^\phi(m_i^\phi, m_0^\phi)$ be distance between the mean of class i and the mean of total and $w(\cdot)$ be a weighted function which is usually a monotonically decreasing function. We define the weighted between-class scatter of the centered samples in the feature space F by

$$\begin{aligned} S_b^{\phi w} &= \tfrac{1}{m}\textstyle\sum_{i=1}^L m_i w(d^\phi(\widetilde{m}_i^\phi, \widetilde{m}_0^\phi)) \\ &\quad (\widetilde{m}_i^\phi - \widetilde{m}_0^\phi)(\widetilde{m}_i^\phi - \widetilde{m}_0^\phi)^T, \end{aligned}$$

where $w(d^\phi(\widetilde{m}_i^\phi, \widetilde{m}_0^\phi)) = (d^\phi(\widetilde{m}_i^\phi, \widetilde{m}_0^\phi))^{-q}$ and $q \geq 2$. If $w(\cdot) = 1$, the matrix $S_b^{\phi w}$ will degenerate to the matrix S_b^ϕ defined in the KDA. In this paper, we use the Euclidean distance. Since $\widetilde{m}_0^\phi = 0$ and $\widetilde{m}_i^\phi = \widetilde{\Phi}(X)U_i$, where $U_i = \frac{1}{m_i}(\underbrace{0, \cdots, 0}_{m_1+\cdots+m_{i-1}}, \underbrace{1, \cdots, 1}_{m_i}, \underbrace{0, \cdots, 0}_{m_{i+1}\cdots+m_L})^T$, we have

$$\begin{aligned} d^\phi(\widetilde{m}_i^\phi, \widetilde{m}_0^\phi) &= \sqrt{(\widetilde{m}_i^\phi)^T(\widetilde{m}_i^\phi)} \\ &= \sqrt{U_i{}^T \widetilde{\Phi}(X)^T \widetilde{\Phi}(X) U_i} \\ &= \sqrt{U_i{}^T \widetilde{K} U_i}. \end{aligned}$$

Putting $\Delta_i = \sqrt{U_i{}^T \widetilde{K} U_i}$, we can deduce that

$$\begin{aligned} S_b^{\phi w} &= \tfrac{1}{m}\textstyle\sum_{i=1}^L m_i w(d^\phi(\widetilde{m}_i^\phi, \widetilde{m}_0^\phi))(\widetilde{m}_i^\phi)(\widetilde{m}_i^\phi)^T \\ &= \frac{1}{m}\sum_{i=1}^L m_i \Delta_i^{-q} \widetilde{\Phi}(X) U_i U_i^T \widetilde{\Phi}(X)^T \\ &= \frac{1}{m}\widetilde{\Phi}(X)\sum_{i=1}^L m_i \Delta_i^{-q} U_i U_i^T \widetilde{\Phi}(X)^T \\ &= \frac{1}{m}\widetilde{\Phi}(X)\sum_{i=1}^L m_i (\Delta_i^{-q/2} U_i)(\Delta_i^{-q/2} U_i)^T \\ &\quad \widetilde{\Phi}(X)^T \\ &= \frac{1}{m}\widetilde{\Phi}(X)\widetilde{W}\widetilde{\Phi}(X)^T, \end{aligned}$$

where $\widetilde{W} = \sum_{i=1}^L m_i (\Delta_i^{-q/2} U_i)(\Delta_i^{-q/2} U_i)^T$. Therefore, the optimal transform matrix V^ϕ can be obtained by maximizing the following Fisher criterion:

$$V^\phi = \arg\max_{V^\phi} \mathrm{tr}(\frac{V^{\phi T} S_b^{\phi w} V^\phi}{V^{\phi T} S_t^\phi V^\phi}). \tag{11}$$

3.2. WKDA/QR algorithm

By means of the kernel trick, the optimization problem (11) can be transformed to the following optimization problem:

$$\widetilde{A} = \arg\max_{\widetilde{A}} \mathrm{tr}(\frac{\widetilde{A}^T SB \widetilde{A}}{\widetilde{A}^T ST \widetilde{A}}), \tag{12}$$

where $V^{\phi} = \widetilde{\Phi}(X)\widetilde{A}, SB = \widetilde{K}\widetilde{W}\widetilde{K} \in R^{m\times m}$ and $ST = \widetilde{K}\widetilde{K} \in R^{m\times m}$. In order to solve the problem (12), we considered two stages: the first stage is to maximize the pseudo between-class scatter matrix SB by QR method and the second stage is to solve a generalized eigenvalue problem. The key problem of the first stage is to deal with the following optimization problem:

$$\hat{A} = \arg\max_{\hat{A}^T\hat{A}=I} \mathrm{tr}(\hat{A}^T SB \hat{A}). \tag{13}$$

Since $\widetilde{W}$ is an $m \times m$ block diagonal symmetric matrix, it is easy to decompose $\widetilde{W}$ into the form $\widetilde{W} = \widetilde{w}\widetilde{w}^T$, where $\widetilde{w} = \mathrm{diag}(\widetilde{w}_1, \cdots, \widetilde{w}_j, \cdots, \widetilde{w}_L)$ is an $m \times L$ matrix and $\widetilde{w}_j$ is a $m_j \times 1$ matrix whose elements are $\Delta_i^{-q/2}/\sqrt{n_j}$. Consequently, $SB = \widetilde{K}\widetilde{w}(\widetilde{K}\widetilde{w})^T = K_1(K_1)^T$, where K_1 is an $m \times L$ matrix.

In general, the number of classes is smaller than the number of training samples. In this case, we can easily prove that $\mathrm{rank(SB)} \leq \mathrm{L}-1$. When L is much smaller than the number of training samples, we can apply QR technique to decompose K_1 and obtain an efficient method for solving kernel discriminant analysis. In fact, if $K_1 = (Q_1\ Q_2)\binom{R}{0}$ is the QR decomposition of K_1, where $R \in R^{r\times L}$ is a row full rank matrix, $r = \mathrm{rank}(SB)$ and $Q_1 \in R^{m\times r}$ and $Q_2 \in R^{m\times(m-r)}$ are volumn orthogonal matrix, we can verify that Q_1 is a solution of the problem (13).

Theorem 1 *For any orthogonal matrix $G \in R^{r\times r}$, $\hat{A} = Q_1 G$ is a solution of the problem (13).*

Proof: Since $G^T G = GG^T = I_r$ and $Q_1^T Q_1 = I_r$, we have $(Q_1 G)^T (Q_1 G) = I_r$ and

$$\begin{aligned}\mathrm{tr}((Q_1G)^T SB(Q_1G)) &= \mathrm{tr}(Q_1^T SBQ_1 GG^T)\\ &= \mathrm{tr}(Q_1^T SBQ_1),\end{aligned}$$

which indicates that the conclusion is true.

Theorem 2 *Let $r = \mathrm{rank}(SB)$ and $K_1 = Q_1 R$ be the QR decomposition of K_1. Let $\widetilde{ST} = Q_1^T STQ_1, \widetilde{SB} = Q_1^T SBQ_1$ and G be a matrix whose columns are the eigenvectors of $(\widetilde{SB})^{-1}\widetilde{ST}$ corresponding to the t largest eigenvalues. Then $Q_1 G$ is an optimal solution of the problem (12).*

Proof: By the QR decomposition of K_1, we know that $\widetilde{SB} = Q_1^T SBQ_1 = R_1 R_1^T$ is nonsingular matrix. According to the definition of the pseudo-inverse of a matrix, we can deduce that

$$\begin{aligned}(SB)^+ &= (K_1(K_1)^T)^+\\ &= (\begin{bmatrix} Q_1\ Q_2 \end{bmatrix}\begin{bmatrix} RR^T & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix} Q_1\ Q_2 \end{bmatrix}^T)^+\\ &= \begin{bmatrix} Q_1\ Q_2 \end{bmatrix}\begin{bmatrix} (RR^T)^{-1} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix} Q_1\ Q_2 \end{bmatrix}^T\end{aligned}$$

and then

$$\begin{aligned}(SB)^+ STg &= (\begin{bmatrix} Q_1\ Q_2 \end{bmatrix}\begin{bmatrix} (RR^T)^{-1} & 0\\ 0 & 0\end{bmatrix}\\ &\quad \begin{bmatrix} Q_1\ Q_2 \end{bmatrix}^T)STg = \lambda g,\end{aligned}$$

which is equivalent to

$$\begin{bmatrix} (RR^T)^{-1} \\ 0 \end{bmatrix} Q_1^T ST \begin{bmatrix} Q_1\ Q_2 \end{bmatrix}\begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} g = \lambda \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} g.$$

Hence,

$$\begin{aligned}(RR^T)^{-1}Q_1^T STQ_1 Q_1^T g &= (\widetilde{SB})^{-1}\widetilde{ST}Q_1^T g\\ &= \lambda Q_1^T g,\end{aligned}$$

which implies that $Q_1^T g$, is a eigenvector of $(\widetilde{SB})^{-1}\widetilde{ST}$ corresponding to the eigenvalue λ. Therefore, the conclusion of the theorem id true.

By Theorem 2, we can propose the following algorithm.

Algorithm 3.1. WKDA/QR

(1) Select a kernel type and compute the kernel matrix K and $\widetilde{K}$;

(2) Calculate matrixes $SB = \widetilde{K}\widetilde{W}\widetilde{K}$ and $ST = \widetilde{K}\widetilde{K}$;

(3) Let $SB = K_1 K_1^T$ and Compute the QR decomposition of K_1: $K_1 = Q_1 R$;

(4) Let $\widetilde{ST} = Q_1^T STQ_1$ and $\widetilde{SB} = Q_1^T SBQ_1$;

(5) Compute the eigenvectors, denoted by G, of the matrix $(\widetilde{SB})^{-1}\widetilde{ST}$ corresponding to the t largest eigenvalues;

(6) Let$\widetilde{A} = Q_1 G$;

(7) For any input vector x, its low dimensional feature representation by WKDA/QR is

$$\begin{aligned}y_x &= \widetilde{A}^T\widetilde{\Phi}(X)^T\phi(x)\\ &= G^T Q_1^T(I - 1_{m\times m}/m)^T(k(x_1, x), \cdots,\\ &\quad k(x_m, x))^T.\end{aligned}$$

3.3. WKDA/SVD algorithm

We know that the problem (11) is equivalent to the following optimization problem (see [16]):

$$W^{\phi} = \arg\max_{W^{\phi}} \mathrm{tr}(\frac{{W^{\phi}}^T S_b^{\phi w} W^{\phi}}{{W^{\phi}}^T S_w^{\phi} W^{\phi}}). \quad (14)$$

With the help of the kernel trick, we can consider the optimization problem:

$$\widetilde{B} = \arg\max_{\widetilde{B}} \mathrm{tr}(\frac{\widetilde{B}^T SB \widetilde{B}}{\widetilde{B}^T SW \widetilde{B}}) \quad (15)$$

and obtain an optimal solution of the problem (14) by $W^{\phi} = \widetilde{\Phi}(X)\widetilde{B}$, where $\widetilde{B}$ is an optimal solution of the problem (15), $SB = \widetilde{K}\widetilde{W}\widetilde{K} \in R^{m\times m}$ and $SW = \widetilde{K}(I - \widetilde{W})\widetilde{K} \in R^{m\times m}$.

Consider the SVD of the matrix SB: $SB = \left[\ U_{b1}\ U_{b2}\ \right] \left[\begin{array}{cc} \Sigma_{b1} & 0 \\ 0 & 0 \end{array}\right] \left[\begin{array}{c} U_{b1}^T \\ U_{b2}^T \end{array}\right]$, where $U_{b1} \in R^{m\times r}$ and $U_{b2} \in R^{m\times(m-r)}$ are column orthogonal matrixes, $\Sigma_{b1} \in R^{r\times r}$ is a diagonal matrix with non-increasing positive diagonal components and $\mathrm{rank}(SB) = r$. It is obvious that the matrix $\widetilde{S}_b = U_{b1}^T SB U_{b1} = \Sigma_{b1}$ is nonsingular. Let $\widetilde{S}_w = U_{b1}^T SW U_{b1}$. In most applications, rank(SW) is greater than rank(SB), since $\mathrm{rank}(U_{b1}^T SW U_{b1}) = \mathrm{rank}(SW) \geq \mathrm{rank}(SB) = \mathrm{rank}(U_{b1}^T SB U_{b1}) = r$. So $\widetilde{S}_w$ is nonsingular (see [28]). We have the following algorithm.

Algorithm 3.2. WKDA/SVD

(1) Select a kernel and compute the kernel matrix K and $\widetilde{K}$;

(2) Let $SB = \widetilde{K}\widetilde{W}\widetilde{K}$ and $SW = \widetilde{K}(I - \widetilde{W})\widetilde{K}$;

(3) Calculate the SVD of SB:

$$SB = \left[\ U_{b1}\ U_{b2}\ \right] \left[\begin{array}{cc} \Sigma_{b1} & 0 \\ 0 & 0 \end{array}\right] \left[\begin{array}{c} U_{b1}^T \\ U_{b2}^T \end{array}\right];$$

(4) Let $\widetilde{S}_w = U_{b1}^T SW U_{b1}$ and $\widetilde{S}_b = U_{b1}^T SB U_{b1}$;

(5) Compute the eigenvectors of the matrix $(\widetilde{S}_w)^{-1}\widetilde{S}_b$, denoted by $\widetilde{G}$, corresponding to the t largest eigenvalues;

(6) Let $\widetilde{B} = U_{b1}\widetilde{G}$;

(7) For any input vector x, its low dimensional feature representation by WKDA/SVD is

$$\begin{aligned} y_x &= \widetilde{B}^T \widetilde{\Phi}(X)^T \phi(x) \\ &= \widetilde{G}^T U_{b1}^T (I - 1_{m\times m}/m)^T (k(x_1, x), \cdots, \\ &\quad k(x_m, x))^T. \end{aligned}$$

4 Experiments and analysis

we evaluate the performance of WKDA/QR and WKDA/SVD algorithms on face recognition task. The publicly available face databases, namely ORL database is used for experiments.

The ORL database contains 40 persons, each having 10 different images. The images of the same person are taken at different times under slightly varying lighting conditions and with various facial experiments. Some people are captured with or without glasses. The heads in images are slightly titled or rotated. The images in the database are manually cropped and recalled to 112×92. In order to reduce the size of the image, we obtain the size of 28×23 pixels. So, the number of features of each character is 644. In the experiments, 8 images are randomly taken from 10 images as training samples sets and the rest are used testing sets. Because training sets are obtained randomly in experiments, there may exist some fluctuation among experiment results. To reduce the fluctuation, we performed each experiment thirty times and all results are an average of them.

All experiments are performed on a PC (2.40 GHZ CPU, 2G RAM) with MATLAB 7.1. Three nonlinear discriminant analysis-based feature extraction methods, namely the proposed WKDA/QR, WKDA/SVD and KDA [11] are tested and compared. For each of the three methods, the face recognition procedure consists of: (1) a feature extraction step where three kinds of feature representation of each training or test sample are extracted by WKDA/QR, WKDA/SVD and KDA [11], respectively, and (2) the nearest neighbor classifier is used.

It is known that proper selection of kernel function is important to achieve better performance in kernel-based learning methods. Generally speaking, there are three classes of widely used kernel functions: polynomial kernels, Gaussian kernels, and sigmoid kernels. To evaluate the effect and stable QR decomposition in WKDA/QR algorithm, we take into consideration polynomial kernels (16) and Gaussian kernels (17):

$$k(x, y) = (x \cdot y + 1)^p. \quad (16)$$

$$k(x, y) = exp(-\|x - y\|^2/2\sigma^2). \quad (17)$$

The parameter p is set as $2, \cdots, 6$, respectively, and the parameter σ is set as 5. we then tested the proposed WKDA/QR, WKDA/SVD and KDA [11] with different parameter p. The weighting function used in WKDA/QR and WKDA/SVD is $w(d^{\phi}) = (d^{\phi})^{-q}(q \geq 2)$. The experiments results are shown in Table 1 and Table 2, respectively.

Table 1: Accuracy rate versus polynomial kernels

P	2	3	4	5	6
WKDA/QR(%)	94.87	94.12	93.21	92.79	92
WKDA/SVD(%)	94.46	93.83	92.17	91.38	90.04
KDA(%)	78.29	77.92	77.75	77	75.75
Feature	39	39	39	39	39

Table 2: Average accuracy rate with $p = 2, 3, 4, 5, 6$ vs. weighted function

q	2	4	6	8
WKDA/QR(%)	95.23	93.50	93.40	93.19
WKDA/SVD(%)	92.60	92.31	92.38	92.42
Feature	39	39	39	39

We can see from Table 1 that WKDA/QR and WKDA/SVD outperform KDA and WKDA/QR outperforms WKDA/SVD based on classification rate under parameter $p = 2, 3, 4, 5, 6$. In addition, the three methods are insensitive to the parameter p of the polynomial kernel function and achieve the best classification with the parameter $p = 2$.

We can see from Table 2 that the weighted function $w(d^{\phi}) = (d^{\phi})^{-q}, (q \geq 2)$ influences the classification accuracy rate of WKDA/QR and WKDA/SVD. To weighted exponent q, WKDA/QR is more insensitive than WKDA/SVD. For different feature extraction tasks, appropriate values of the weighted exponent q should be determined by an experiment with available training set. We examined the classification accuracy rate of WKDA/QR and WKDA/SVD methods with $q = 2, 4, 6, 8$, respectively. In addition, to average classification accuracy rate, WKDA/QR outperforms WKDA/SVD with polynomial kernels.

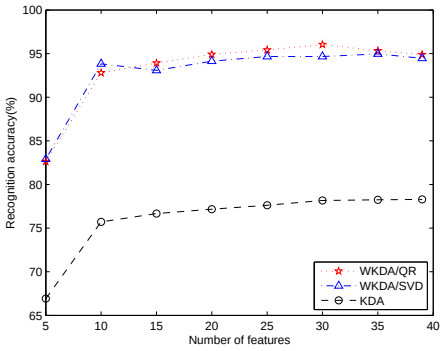

Fig.1(a) accuracy rate curves with 2-polynomial kernel

In order to evaluate QR decomposition in WKDA/QR outperforms SVD in WKDA/SVD on classification accuracy rate, we further compare the two methods on the ORL database. The 2-polynomial kernel (that is $q = 2$) and weighted function $w(d^{\phi}) = (d^{\phi})^{-6}$ are used in the experiments. The experiment results are shown in Fig.1 (a) and Fig.1 (b). Fig.1(a) shows the accuracy rate curves of WKDA/QR, WKDA/SVD and KDA with respect to the feature dimen-

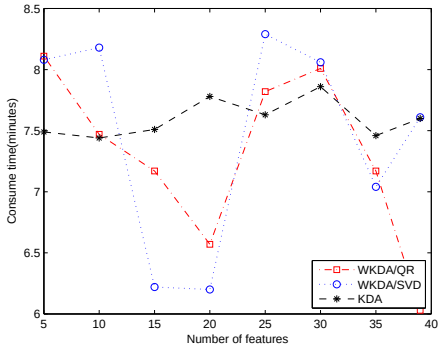

Fig.1(b) time-consuming curves with 2-polynomial kernel

sionality. We can see from Fig.1(a) that, to classification accuracy rate, WKDA/QR and WKDA/SVD are better than KDA, and WKDA/QR is better than WKDA/SVD when feature dimensionality is larger than 10. Fig.1(b) shows the time-consuming curves of WKDA/QR, WKDA/SVD and KDA methods with respect to the feature dimensionality. We can see from Fig.1(b) that, time-consuming, WKDA/QR is more stable than WKDA/SVD and WKDA/QR is better than KDA for high-dimension data.

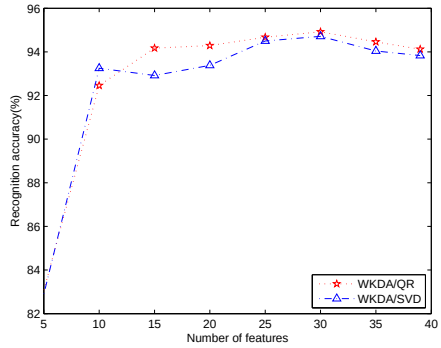

Fig.1(c) accuracy rate curves with 3-polynomial kernel

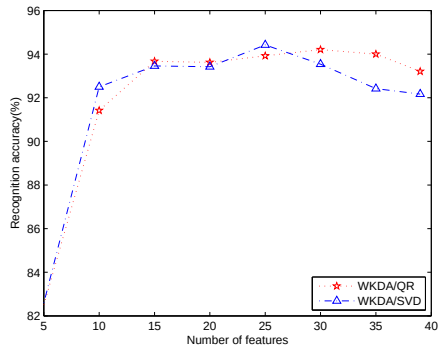

Fig.1(d)) accuracy rate curves with 4-polynomial kernel

In order to evaluate QR decomposition in WKDA/QR outperforms SVD in WKDA/SVD on the in-

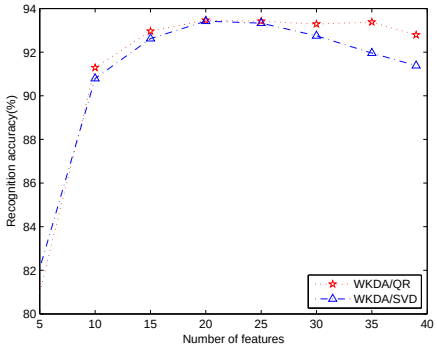

Fig.1(e) accuracy rate curves with 5-polynomial kernel

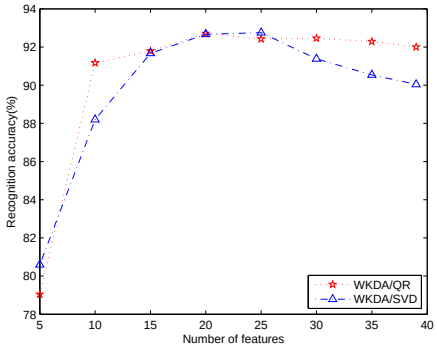

Fig.1(f) accuracy rate curves with 6-polynomial kernel

fluence of the parameters p of polynomial kernels function, we compare the two methods on the ORL database. The parameter p is taken as 3,4,5 and 6, respectively, and the weighted function $w(d^{\phi}) = (d^{\phi})^{-6}$ is used in the experiments. The experiments results are shown in Fig.1(c), Fig.1(d), Fig.1(e), and Fig.1(f), respectively. We can see from Fig.1(c), Fig.1(d), Fig.1(e) and Fig.1(f) that, under different polynomial kernels function, WKDA/QR outperforms WKDA/SVD when the number of feature dimensionality is more than 10.

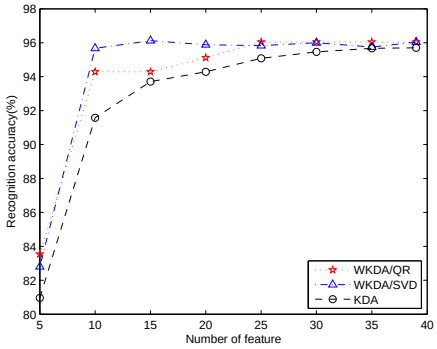

Fig.2(a) accuracy rate curves with $\sigma = 5$ and $q = 6$

In order to improve the classification accuracy rate and evaluate the effective of QR decomposition,

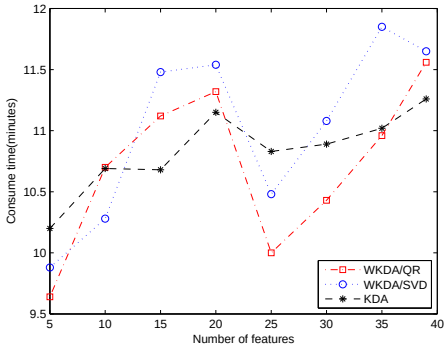

Fig.2(b) time-consuming versus with $\sigma = 5$ and $q = 6$

we use Gaussian kernel with $\sigma = 5$ and the weighted function $w(d^{\phi}) = (d^{\phi})^{-6}$ in experiments. Fig.2(a) and Fig.2(b) show the accuracy rate curves and time-consuming curves of WKDA/QR, WKDA/SVD and KDA with respect to the feature dimensionality, respectively. We can see from Fig.2(a) that, on classification accuracy rate, WKDA/QR and WKDA/SVD outperform KDA, and WKDA/QR is more effective and stable than WKDA/SVD when the feature dimensionality is larger than 25. We can see from Fig.2(b) that, to time-consuming, WKDA/QR is more stable than WKDA/SVD, and WKDA/QR is better than KDA for high-dimension data.

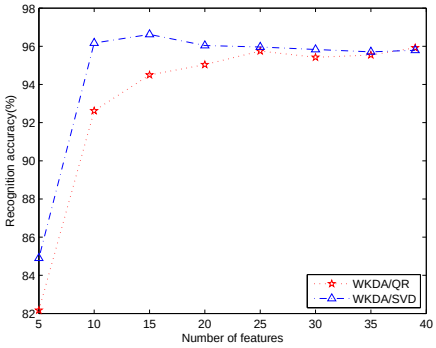

Fig.2(c) accuracy rate curves with $\sigma = 5$ and $q = 2$

In the following, we consider the influence of the weighted exponent q and the parameter σ of the Gaussian kernel function. We first consider the influence of the weighted exponent q. We compare the classification accuracy rate of WKDA/QR and WKDA/SVD on the ORL database with $q = 2, 3, 4, 8$, respectively, in which Gaussian kernel function with $\sigma = 5$ is used. The experiment results are shown in Fig.2(c), Fig.2(d), Fig.2(e) and Fig.2(f), respectively. From them, we can see that WKDA/SVD outperforms WKDA/QR, which indicates that the selection of the parameter q of weighted function $w(d^{\phi}) = (d^{\phi})^{-q}$ is very important for classification accuracy rate.

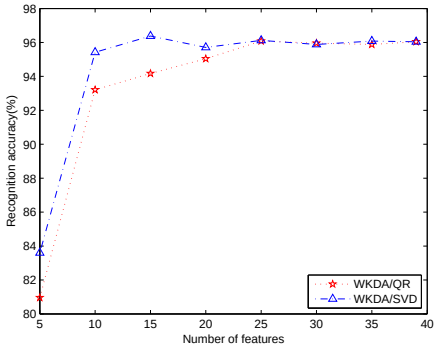

Fig.2(d) accuracy rate curves with $\sigma = 5$ and $q = 3$

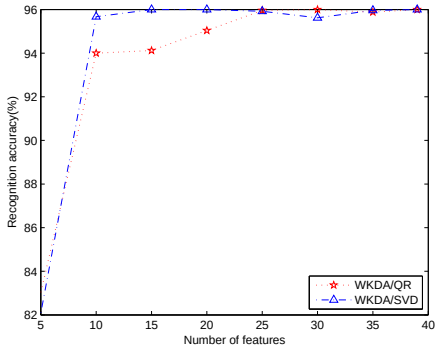

Fig.2(f) accuracy rate curves with $\sigma = 5$ and $q = 8$

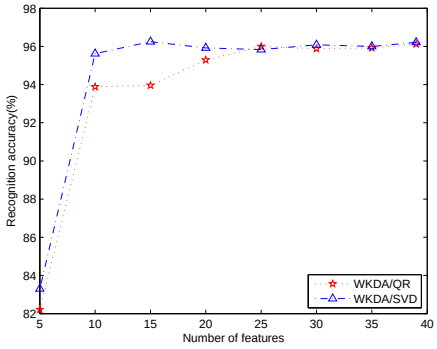

Fig.2(e) accuracy rate curves with $\sigma = 5$ and $q = 4$

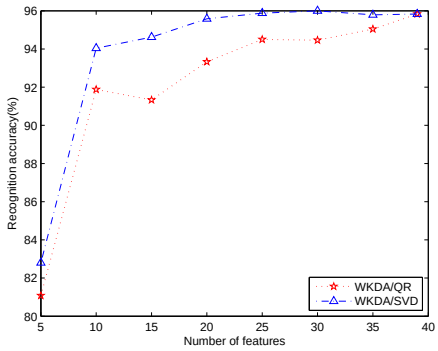

Fig.3(a) accuracy rate curves with $q = 6$ and $\sigma = 3.5$

Next, we consider the influence of the parameter σ of the Gaussian kernel function. We compare the classification accuracy rate of WKDA/QR and WKDA/SVD on the ORL database with $\sigma = 3.5, 4$ and 4.5, respectively, in which weighted function with $q = 6$ is used. The experiment results are shown in Fig.3(a), Fig.3(b) and Fig.3(c), respectively. From them, we can see that WKDA/SVD outperforms WKDA/QR, which indicates that the selection of the parameter σ of the Gaussian kernel function also is very important for classification accuracy rate. However, with the feature dimensionality is 39, WKDA/QR and WKDA/SVD are the same.

5 Conclusion

In this paper, we present two kinds of kernel-based weighted discriminant analysis (WKDA) methods, WKDA/QR and WKDA/SVD methods, for feature extraction with combination of a weighted scheme and QR decomposition and singular value decomposition technique. The two methods can find lower-dimensional nonlinear features with significant discriminant power and can be viewed as a generalization of KDA. Experiments show that QR decomposition is an efficient and effective step which can save much time for high-dimensional database, and then WKDA/QR algorithm is effective and feasible in real world application.

In order to compare WKDA/SVD and WKDA/QR methods, we select different parameters of kernel functions and weighted function. Experiments results show that the selection of parameters is very important for classification accuracy rate. For Gaussian kernel function, WKDA/QR is more sensitive than WKDA/SVD on classification accuracy rate.

Acknowledgements: The research is supported by NSF (grant No.10871226), NSF (grant No. ZR2009AL006) of Shandong Province and Young and Middle-Aged Scientists Research Foundation (grant No. BS2010SF004) of Shandong Province, P.R. China. Corresponding author: Liya Fan, fanliya63@126.com.

References:

[1] J. Duchene and S. Leclercq, An optimal transformation for dimension and principal component analysis, *IEEE Trans, PAMI.* 10 (6), 1988, pp. 978–983.

[2] D.H. Foley, J.W. Sammon Jr., An optimal set of discriminant vectors, *IEEE Trans, Comput.* 24 (3), 1975, pp. 281–289.

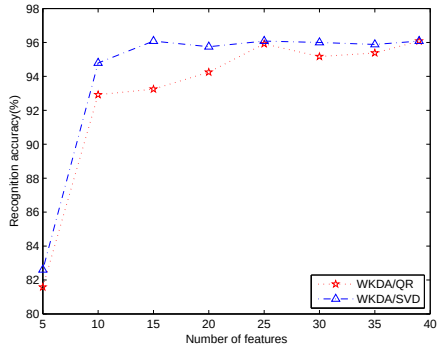

Fig.3(b) accuracy rate curves with $q = 6$ and $\sigma = 4$

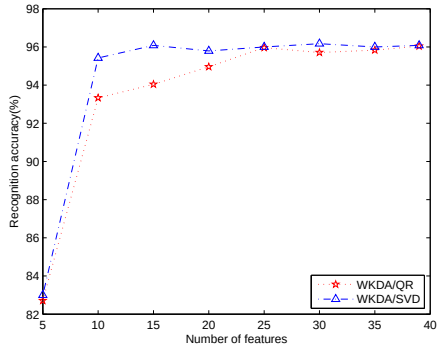

Fig.3(c) accuracy rate curves with $q = 6$ and $\sigma = 4.5$

[3] P. Belhumeur, J. Hespanha, D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection, *IEEE Trans Pattern Anal Mach Intell.* 19, 1997, pp. 711–720.

[4] Martnez AM, Kak AC, PCA versus LDA, *IEEE Trans Pattern Anal Mach Intell.* 23, 2001, pp. 228–233.

[5] J. Ye, Q. Li, LDA/QR: An efficient and effective dimension reduction algorithm and its theoretical foundation, *Pattern Recog.* 37, 2004, pp. 851–854.

[6] Z. Jin, J.Y. Yang, Z.S. Hu, Z. Luo, Face recognition based on the uncorrelated discriminant transformation, *Pattern Recog.* 34 (7), 2001, pp. 1405–1416.

[7] Z. Jin, J.Y. Yang, Z. Tang, Z. Hu, A theorem on the uncorrelated optimal discriminant vectors, *Pattern Recog.* 34 (10), 2001, pp. 2041–2047.

[8] Jing Xiao-Yuan, Zhang David, Jin Zhong, Improvements on the uncorrelated optimal discriminant vectors, *Pattern Recog.* 36 (8), 2003, pp. 1921–1923.

[9] B. Scholkopf, A. Smola, K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Comput.* 10 (5), 1998, pp. 1299–1319.

[10] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, Muller KR Fisher Discriminant Analysis with Kernels. In: *Proceedings of IEEE international workshop neural networks for signal processing IX.* 1999, pp. 41–48.

[11] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, *Neural Comput.* 12 (10), 2000, pp. 2385–2404.

[12] Yang MH, Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods. In: Proceedings of fifth IEEE international conference automatic face and gesture recognition, *IEEE Press, New York.* 2002, pp. 215–220.

[13] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, Muller KR Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces, *IEEE Trans Pattern Anal Mach Intell.* 25, 2003, pp. 623–628.

[14] Lu J, Plataniotis KN, Venetsanopoulos AN, Face recognition using kernel direct discriminant analysis algorithms, *IEEE Trans on Neural Netw.* 14, 2003, pp. 117–126.

[15] Jian Y., Z. Jin, D. Zhang, The essence of kernel Fisher discriminant: KPCA plus LDA, *Pattern Recogn.* 37, 2004, pp. 2097–2100.

[16] Jian Y., A.F. Frangi, J.Y. Yang, D. Zhang, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, *IEEE Trans Pattern Anal Mach Intell.* 27, 2005, pp. 230–244.

[17] Lotlikar R., Kothari R., Fractional-step dimension reduction, *IEEE Trans Pattern Anal Mach Intell.* 22, 2000, pp. 623–627.

[18] Loog M., Duin R.P.W., Haeb-Umbach R., Multiclass linear dimension reduction by weighted pairwise Fisher criteria, *IEEE Trans Pattern Anal Mach Intell.* 23, 2001, pp. 762–766.

[19] Dai G., Qian Y.T., Jia S., A kernel fractional-step nonlinear discriminant analysis for pattern recognition. In: *Proceedings of the 18th international conference on pattern recognition.* 2004, pp. 431–434.

[20] Zhou D., Yang X., Peng N., Improved-LDA based face recognition using both facial global and local information, *Pattern Recogn. Lett.* 27, 2006, pp. 537–543.

[21] H. Yu, J. yang, A direct LDA algorithm for high-dimensional data with application to face recognition, *Pattern Recogn.* 34, 2001, pp. 2067–2070.

[22] Dai G., Yeung D.Y., Qian Y.T. Face recognition using a kernel fractional-step discriminant analysis algorithm, *Pattern Recogn.* 40, 2007, pp. 22–243.

[23] Zhou D., Tang Z., Kernel-based improved discriminant analysis and its application to face recognition, *soft comput.* 14, 2010, pp. 103–111.

[24] Zhou D., Tang Z., A modification of kernel discriminant analysis for high-dimensional data-with application to face recognition, *signal processing.* 90, 2010, pp. 2423–2430.

[25] K. Fukunaga, Introduction to statistical pattern classification, *academic press, San Diego, California, USA*. 1990.

[26] Scholkopf B., Smola A., Moller K.R., Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computat.* 10, 1999, pp. 1299–1319.

[27] Jieping Ye, Qi Li, LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation, *Pattern Recog.* 37 (4), 2004, pp. 851–854.

[28] C.H. Park, H. Park, A comparison of generalized linear discriminant analysis algorithms, *Pattern Recog.* 41, 2008, pp. 1083–1097.