Int. J. Appl. Math. Comput. Sci., 2013, Vol. 23, No. 4, 887–903
DOI: 10.2478/amcs-2013-0066

A PRACTICAL APPLICATION OF KERNEL–BASED FUZZY DISCRIMINANT ANALYSIS

JIAN-QIANG GAO *, LI-YA FAN **, LI LI ***, LI-ZHONG XU *

* College of Computer and Information Engineering
Hohai University, Nanjing, 210098, PR China
e-mail: jianqianggaoHH@126.com,jianqianggaohhedu@gmail.com,gaojq@yahoo.in

**School of Mathematics Sciences
Liaocheng University, Shandong, 252059, PR China
e-mail: fanliya63@126.com

***Department of Mathematics
Nanjing University of Finance and Economics, Nanjing, 210023, PR China
e-mail: lili880827@126.com

A novel method for feature extraction and recognition called Kernel Fuzzy Discriminant Analysis (KFDA) is proposed in this paper to deal with recognition problems, e.g., for images. The KFDA method is obtained by combining the advantages of fuzzy methods and a kernel trick. Based on the orthogonal-triangular decomposition of a matrix and Singular Value Decomposition (SVD), two different variants, KFDA/QR and KFDA/SVD, of KFDA are obtained. In the proposed method, the membership degree is incorporated into the definition of between-class and within-class scatter matrices to get fuzzy between-class and within-class scatter matrices. The membership degree is obtained by combining the measures of features of samples data. In addition, the effects of employing different measures is investigated from a pure mathematical point of view, and the t-test statistical method is used for comparing the robustness of the learning algorithm. Experimental results on ORL and FERET face databases show that KFDA/QR and KFDA/SVD are more effective and feasible than Fuzzy Discriminant Analysis (FDA) and Kernel Discriminant Analysis (KDA) in terms of the mean correct recognition rate.

Keywords: kernel fuzzy discriminant analysis, fuzzy k-nearest neighbor, QR decomposition, SVD, fuzzy membership matrix, t-test.

1. Introduction

Face recognition has been investigated in various areas such as pattern recognition, image processing, and computer vision. In practice, face recognition is a very difficult problem due to a substantial variation in light direction, different face poses, and diversified facial expressions. Linear Discriminant Analysis (LDA) is a well-known supervised classifier in statistical pattern recognition, which is widely used as a dimensionality reduction technique in face recognition, but it cannot be applied directly to small sample problems (Raudys and Jain, 1991) due to the singularity of the within-class scatter matrix. In order to use LDA for small sample problems such as face recognition, much research has been done (e.g., Gao *et al.*,2008; Koc and Barkana, 2011; Duda *et al.*, 2012; Fukunaga, 1990; Hastie *et al.*, 1991; 1994; 1995; Liu *et al.*, 2008; Jain and Zongker, 1997; Lee *et al.*, 2001; Pal and Eluri, 1998; Swets and Weng, 1996; Belhumeur and Kriegman, 1997; Yang and Yang, 2003; 2001; Hong and Yang, 2005; Friedman, 1989).

The most popular approach, the Fisher face, was proposed by Swets and Weng (1996) as well as Belhumeur and Kriegman (1997). There, Principal Component Analysis (PCA) is first used to reduce the dimension of the original space and then the classical Fisher Linear Discriminant Analysis (FLDA) is applied to reduce the space dimension. A limitation of the Fisher face is that some effective discriminatory information may be lost and the PCA step cannot guarantee the transformed within-class scatter matrix to be nonsingular. Woźniak and Krawczyk (2012) present a significant modification to

the AdaSS (Adaptive Splitting and Selection) algorithm. The method is based on simultaneous partitioning the feature space and an assignment of a compound classifier to each of the subsets.

In order to deal with the singularity problem, a popular method is to add a singular value perturbation to the within-class scatter matrix (e.g., Hong and Yang, 2005). Penalized Discriminant Analysis (PDA) is another regularized method (e.g., Hastie *et al.*, 1994; 1995). Its goals are not only to overcome small sample problems but also to smooth the coefficients of discriminant vectors. The methods based on the null subspace have LDA+PCA and direct-LDA (e.g., Chen *et al.*, 2000; Yu and Yang, 2001). Zhuang and Dai (2005; 2007) develop an Inverse Fisher Discriminant Analysis (IFDA) method, which modifies the procedure of PCA and derives regular and irregular information from the within-class scatter matrix by the inverse Fisher discriminant criterion.

Recently, many kernel-based algorithms have been proposed, such as Support Vector Machines (SVMs) (Vapnik, 1998), Kernel Fisher Discriminant Analysis (KFDA), Kernel Principal Component Analysis (KPCA) (Schölkopf *et al.*, 1998), Kernel Canonical Correlation Analysis (KCCA) (Liu and Xue, 2012), kernel fuzzy Support Vector Regressions (SVRs) (Loog *et al.*, 2001), particle swarm optimization Kernel-based Principal Component Analysis (KPCA) and support vector machines for an electric Power Quality (PQ) problem classification (Pahasa and Ngamroo, 2012), Weighted Kernel Discriminant Analysis (WKDA) (Gao and Fan, 2011), or the range space of the between-class scatter matrix principal component analysis method (PCA/range(S_b)) (Gao *et al.*, 2012). We can also mention here the within-class scatter matrix null space median method (M-N(S_w)) (Gao *et al.*, 2013).

Świercz (2010) proposed a classification algorithm which is based on the matching shape idea of non-stationary signals available from observations. By taking advantage of the technology of fuzzy sets (Zadeh, 1965), some studies have been carried out for fuzzy pattern recognition (e.g., Kwak and Pedrycz, 2005; Keller *et al.*, 1985; Zheng *et al.*, 2006b; 2005a; Wu and Zhou, 2006; Yang *et al.*, 2009). Zheng *et al.* (2006b) proposed a kernel Fisher discriminant algorithm with fuzzy set theory (FKFD). The key idea of FKFD is that KPCA transformation is implemented in the original image space to transform all samples into a low-dimension space with a kernel trick, and then the FKNN algorithm is implemented in the KPCA transformed space.

In this paper, inspired by the above works, we extend Fuzzy Discriminant Analysis (FDA) to a nonlinear model and obtain a new learning method called kernel fuzzy discriminant analysis. The main idea of KFDA is that the measure computation and the fuzzy membership matrix U are implemented in the original image space with the help of the FKNN algorithm, and then the kernel transformation is implemented with a kernel trick and a fuzzy membership matrix U. A key step of measure is how to incorporate the contribution of each training sample into the fuzzy membership matrix U. Detailed descriptions will be displayed in the following. Meanwhile, based on QR decomposition and SVD, we get two different variants, KFDA/QR and KFDA/SVD, of KFDA. Since QR decomposition on a small size matrix is adopted, two superiorities of our method are their computational efficiency and their ability of avoiding singularities. In the proposed method, the membership degree is incorporated into the definition of between-class and within-class scatter matrices to get fuzzy between-class and within-class scatter matrices. According to the recognition rates, we compare our method with FDA/QR, FDA/SVD, KDA/QR and KDA/SVD under different measures and kernel functions. Experimental results on ORL and FERET face databases show that KFDA compares favorably with FDA and KDA.

The rest of this paper is organized as follows. Linear discriminant analysis, Kernel Discriminant Analysis (KDA) and fuzzy discriminant analysis are briefly introduced and discussed in Section 2. Detailed descriptions of KFDA/QR, KFDA/SVD and different measures are produced in Section 3. In Section 4, in order to demonstrate the efficiency of the method we proposed, many experiments are done under different measures. Conclusions and future work are summarized in Section 5.

2. Review of LDA, KDA and FDA

2.1. LDA. In this subsection, we first introduce some notation. Given a data matrix $X = [x_1, \ldots, x_N] \in \mathbb{R}^{n\times N}$, where $x_1, \ldots, x_N \in \mathbb{R}^n$ are samples, we consider finding a linear transformation $G \in \mathbb{R}^{n\times l}$ that maps each x_i to $y_i \in \mathbb{R}^l$ with $y_i = G^T x_i$. Assume that the original data in X are partitioned into c classes as $X = [X_1, \ldots, X_c]$, where $X_i \in \mathbb{R}^{n\times n_i}$ contains data points of the i-th class and $\sum_{i=1}^{c} n_i = N$. In discriminant analysis, between-class, within-class and total scatter matrices are respectively defined as follows (Fukunaga, 1990):

$$\begin{aligned} S_b &= \frac{1}{N}\sum_{i=1}^{c} n_i (m_i - m_0)(m_i - m_0)^T, \\ S_w &= \frac{1}{N}\sum_{i=1}^{c}\sum_{x\in X_i} (x - m_i)(x - m_i)^T, \\ S_t &= \frac{1}{N}\sum_{i=1}^{N} (x_i - m_0)(x_i - m_0)^T, \end{aligned} \tag{1}$$

where $m_i = (1/n_i)\sum_{j=1}^{n_i} x_j$ is the centroid of the i-th class and $m_0 = (1/N)\sum_{j=1}^{N} x_j$ is the global centroid of the training data set.

LDA aims to find an optimal transformation G such that the class structure of the original high-dimensional space is preserved in the low-dimensional space. From (1), we can easily show that $S_t = S_b + S_w$ and see that the traces

$$(S_b) = \frac{1}{N}\sum_{i=1}^{c} n_i \|m_i - m_0\|_2^2$$

and

$$(S_w) = \frac{1}{N}\sum_{i=1}^{c}\sum_{x\in X_i} \|x - m_i\|_2^2$$

measure the closeness of vectors within the classes and the separation between classes, respectively.

In the low-dimensional space resulting from the linear transformation G, the between-class, within-class and total scatter matrices become $S_b^L = G^T S_b G$, $S_w^L = G^T S_w G$ and $S_t^L = G^T S_t G$, respectively. An optimal transformation G would maximize trace(S_b^L) and minimize trace(S_w^L). Common optimization problems in LDA include (see Fukunaga, 1990)

$$\max_G \mathrm{tr}\{(S_w^L)^{-1} S_b^L\} \text{ and } \min_G \mathrm{tr}\{(S_b^L)^{-1} S_w^L\}. \quad (2)$$

The optimization problems in (2) are equivalent to finding generalized eigenvectors satisfying $S_b g = \lambda S_w g$ with $\lambda \neq 0$. The solution can be obtained by using the eigen-decomposition to the matrix $S_w^{-1} S_b$ if S_w is nonsingular or $S_b^{-1} S_w$ if S_b is nonsingular. It was shown by Fukunaga (1990) that the solution can also be obtained by computing the eigen-decomposition of the matrix $S_t^{-1} S_b$ if S_t is nonsingular. There are at most $c-1$ eigenvectors corresponding to nonzero eigenvalues since the rank of the matrix S_b is bounded from above by $c-1$. Therefore, the number of retained dimensions in LDA is at most $c-1$. A stable way to compute the eigen-decomposition is to apply SVD on the scatter matrices. Details can be found in the work of Swets and Weng (1996).

2.2. KDA. KDA is a kernel version of LDA to deal with feature extraction and classification of nonlinear characteristics. The basic idea of KDA is to firstly project original patterns into a high-dimensional feature space $\mathcal{F}$ by an implicit nonlinear mapping $\phi : \mathbb{R}^n \to \mathcal{F}$ and then to the use LDA in the feature space $\mathcal{F}$.

Let us consider a set of N training samples $\{x_1, x_2, \ldots, x_N\}$ taking values in an n-dimensional space. Let c be the number of classes and n_i be the number of training samples in the i-th class, $i = 1, \ldots, c$. Obviously, $N = \sum_{i=1}^{c} n_i$. In general, the Fisher criterion (Fukunaga,1990; Zheng *et al.*, 2006b) can be defined as

$$\max_w J(w) = \frac{w^T S_b^\phi w}{w^T S_t^\phi w}, \quad (3)$$

where $S_b^\phi = \frac{1}{N}\sum_{i=1}^{c} n_i (m_i^\phi - m_0^\phi)(m_i^\phi - m_0^\phi)^T$ and $S_t^\phi = \frac{1}{N}\sum_{i=1}^{N} (\phi(x_i) - m_0^\phi)(\phi(x_i) - m_0^\phi)^T$ are the between-class and total scatter matrices defined in the feature space $\mathcal{F}$, 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 (3) can be transformed into the following eigenvalue problem:

$$S_b^\phi w = \lambda S_t^\phi w. \quad (4)$$

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

$$\begin{aligned}\widetilde{K} &= \widetilde{\Phi}(X)^T \widetilde{\Phi}(X) \\ &= (I - 1_{N\times N}/N)^T K (I - 1_{N\times N}/N).\end{aligned} \quad (5)$$

According to reproducing kernel theory (Schölkopf *et al.*, 1998), the eigenvector w lies in the span of $\{\widetilde{\phi}(x_1), \ldots, \widetilde{\phi}(x_N)\}$ and then there exist coefficients $b_i, i = 1, 2, \ldots, N$, such that

$$w = \sum_{i=1}^{N} b_i \widetilde{\phi}(x_i) = \widetilde{\Phi}(X) b, \quad (6)$$

where $b = (b_1, \ldots, b_N)^T$.

Let $W = \mathrm{diag}(s_1, \ldots, s_j, \ldots, s_c)$, where s_j is an $n_j \times n_j$ matrix whose elements are $1/n_j$. Substituting (6) into (3), we obtain

$$\max_b J(b) = \frac{b^T \widetilde{K} W \widetilde{K} b}{b^T \widetilde{K}\widetilde{K} b}. \quad (7)$$

In general, the vector b_1 corresponding to the maximal value of $J(b)$ is an 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 $b_1, \ldots, b_t$ are t optimal discriminant directions and $B = [b_1, b_2, \ldots, b_t]$. Then B should satisfy

$$B = \arg\max_B \mathrm{tr}\Big(\frac{B^T S_b^* B}{B^T S_t^* B}\Big), \quad (8)$$

where $S_b^* = \widetilde{K}W\widetilde{K}$, $S_t^* = \widetilde{K}\widetilde{K}$. The optimization problem (8) can be transformed into the following generalized eigenvalue problem:

$$S_b^* a = \lambda S_t^* a. \tag{9}$$

The solution to the problem (9) can be obtained by solving the generalized eigenvalue problem. Suppose that $\lambda_1, \lambda_2, \ldots, \lambda_t$ are the t largest eigenvalues of the problem (9) sorted in descending order and $b_1, \ldots, b_t$ are the corresponding eigenvectors. We can obtain the KDA transform matrix by

$$W = [w_1, \ldots, w_t] = \widetilde{\Phi}(X)[b_1, \ldots, b_t] = \widetilde{\Phi}(X)B. \tag{10}$$

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

$$\begin{aligned} y_x &= W^T\widetilde{\phi}(x) \\ &= B^T\widetilde{\Phi}(X)^T\widetilde{\phi}(x) \\ &= A^T(\widetilde{k}(x_1, x), \widetilde{k}(x_2, x), \ldots, \widetilde{k}(x_N, x))^T. \end{aligned} \tag{11}$$

2.3. FDA. Kwak and Pedrycz (2005) proposed the fuzzy Fisher face method for recognition via fuzzy sets. A fuzzy c class partition of these vectors specifies the degree of membership of each vector to the classes. The membership matrix $U = [u_{ij}]\,(i = 1, 2, \ldots, c, j = 1, 2, \ldots, N)$ can be obtained by the Fuzzy k-Nearest Neighbor (FKNN) (Keller *et al.*, 1985). The FKNN will be discussed in Section 3.1. Taking into account the membership grades, the mean vector of each class $\widetilde{m}_i$ is calculated as follows:

$$\widetilde{m}_i = \frac{\sum_{j=1}^{N} u_{ij}x_j}{\sum_{j=1}^{N} u_{ij}}. \tag{12}$$

The between-class fuzzy scatter matrix S_{Fb} and the within-class fuzzy scatter matrix S_{Fw} incorporate the membership values in their calculation

$$S_{Fb} = \sum_{i=1}^{c} n_i(\widetilde{m}_i - m_0)(\widetilde{m}_i - m_0)^T, \tag{13}$$

$$S_{Fw} = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_i^j - \widetilde{m}_i)(x_i^j - \widetilde{m}_i)^T. \tag{14}$$

The optimal fuzzy projection matrix G of the fuzzy Fisher face follows the expression

$$G = \arg\max_G \frac{|G^T S_{Fb} G|}{|G^T S_{Fw} G|}. \tag{15}$$

Finally, PCA plus fuzzy LDA are used in small size sample cases.

3. Concept of kernel fuzzy discriminant analysis

LDA and FDA are linear learning algorithms and they cannot deal with nonlinear problems. KDA is a kernel version of LDA which deals with feature extraction and classification of nonlinear characteristics. The basic idea of KDA can be achieved mainly via reproducing kernel theory.

The main idea of KFDA is that measure computation and the fuzzy membership matrix U are implemented in the original images space with the help of the FKNN algorithm, and then the kernel transformation is implemented with the kernel trick and fuzzy membership matrix U. In the second step, the original patterns are projected from the input space $\mathbb{R}^n$ into the feature space $\mathcal{F}$ by an implicit nonlinear mapping $\phi : \mathbb{R}^n \to \mathcal{F} : x \to \Phi(x)$. We do not need to calculate the mapping in the high-dimension feature space because the kernel function can do it implicitly. Then the dimension disaster problem can be avoided. The heart of the matter is how to incorporate the contribution of each training sample into the fuzzy membership matrix U with the help of a measure. In this paper, the fuzzy membership matrix U is determined via a measure of all features of each training sample. That is to say, the measure controls the influence of the fuzzy membership degree. In this method, the fuzzy membership degree is incorporated into the definition of between-class and within-class scatter matrices to get fuzzy between-class and within-class scatter matrices. The fuzzy membership degree and each class center are obtained with the FKNN algorithm (Keller *et al.*, 1985).

In general, in the higher dimension feature space $\mathcal{F}$, LDA can be achieved by maximizing the following Fisher criterion (e.g., Fukunaga, 1990; Yang *et al.*, 2005; Schölkopf *et al.*, 1998):

$$\max_v J(v) = \frac{v^T S_{Fb}^{\phi} v}{v^T S_{Ft}^{\phi} v}, \tag{16}$$

where $S_{Fb}^{\phi} = \frac{1}{N}\sum_{i=1}^{c} n_i(\widetilde{m}_i^{\phi} - m_0^{\phi})(\widetilde{m}_i^{\phi} - m_0^{\phi})^T$ and $S_{Ft}^{\phi} = \frac{1}{N}\sum_{j=1}^{N}(\phi(x_j) - m_0^{\phi})(\phi(x_j) - m_0^{\phi})^T$ are the between-class fuzzy and total fuzzy scatter matrices defined in the feature space $\mathcal{F}$, respectively. Here $\widetilde{m}_i^{\phi}$ 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 (16) can be transformed into the following eigenvalue problem:

$$S_{Fb}^{\phi} v = \lambda S_{Ft}^{\phi} v. \tag{17}$$

Let $a = (a_1, a_2, a_3, \ldots, a_{N-1}, a_N)^T$, $M = \mathrm{diag}(n_1, \ldots, n_j, \ldots, n_c)$, where n_j is the number of training samples in the j-th class. We define the between-class fuzzy scatter and total fuzzy scatter

matrices of the centered samples in the feature space $\mathcal{F}$, respectively,

$$\begin{aligned} S_{Fb}^{\phi} &= \frac{1}{N}\sum_{i=1}^{c} n_i(\widetilde{m}_i^{\phi} - \overline{m}^{\phi})(\widetilde{m}_i^{\phi} - \overline{m}^{\phi})^T \\ &= \frac{1}{N}\sum_{i=1}^{c} n_i(\widetilde{\Phi}(X)U^T e_i)(\widetilde{\Phi}(X)U^T e_i)^T \\ &= \frac{1}{N}\widetilde{\Phi}(X)U^T\sum_{i=1}^{c} n_i e_i e_i^T U\widetilde{\Phi}(X)^T \\ &= \frac{1}{N}\widetilde{\Phi}(X)U^T MU\widetilde{\Phi}(X)^T, \end{aligned} \tag{18}$$

$$\begin{aligned} S_{Ft}^{\phi} &= \frac{1}{N}\sum_{j=1}^{N}(\phi(x_j) - \overline{m}^{\phi})(\phi(x_j) - \overline{m}^{\phi})^T \\ &= \frac{1}{N}\sum_{j=1}^{N}(\widetilde{\Phi}(X)E_j)(\widetilde{\Phi}(X)E_j)^T \\ &= \frac{1}{N}\widetilde{\Phi}(X)\sum_{j=1}^{N} E_j E_j^T\widetilde{\Phi}(X)^T \\ &= \frac{1}{N}\widetilde{\Phi}(X)E\widetilde{\Phi}(X)^T \\ &= \frac{1}{N}\widetilde{\Phi}(X)\widetilde{\Phi}(X)^T, \end{aligned} \tag{19}$$

where

$$e_i = [\underbrace{0,\ldots,0}_{0+\cdots+(i-1)}, \underbrace{1}_{i}, \underbrace{0,\ldots,0}_{c-i}]^T,$$

$$E_j = [\underbrace{0,\ldots,0}_{0+\cdots+(j-1)}, \underbrace{1}_{j}, \underbrace{0,\ldots,0}_{N-j}]^T.$$

Therefore, we can obtain the following equation:

$$\max_a J(a) = \frac{a^T\widetilde{K}U^T MU\widetilde{K}a}{a^T\widetilde{K}\widetilde{K}a}. \tag{20}$$

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 for feature extraction. Hence, it is often necessary to obtain t ($t > 1$) optimal discriminant directions. Assume that $a_1, \ldots, a_t$ are t optimal discriminant directions and $A = [a_1, a_2, \ldots, a_t]$. Then A should satisfy

$$A = \arg\max_A \operatorname{tr}\Big(\frac{A^T S_b' A}{A^T S_t' A}\Big), \tag{21}$$

where $S_b' = \widetilde{K}U^T MU\widetilde{K}$, $S_t' = \widetilde{K}\widetilde{K}$. The optimization problem (21) can be transformed into the following generalized eigenvalue problems:

$$S_b' a = \lambda S_t' a. \tag{22}$$

The solution of the problem (22) can be obtained by solving the generalized eigenvalue problem. Suppose that $\lambda_1, \lambda_2, \ldots, \lambda_t$ are the t largest eigenvalues of the problem (22) sorted in descending order and $a_1, \ldots, a_t$ are the corresponding eigenvectors. We can obtain the KFDA transform matrix by

$$V = [v_1, \ldots, v_t] = \widetilde{\Phi}(X)[a_1, \ldots, a_t] = \widetilde{\Phi}(X)A. \tag{23}$$

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

$$\begin{aligned} y_x &= V^T\widetilde{\phi}(x) \\ &= A^T\widetilde{\Phi}(X)^T\widetilde{\phi}(x) \\ &= A^T(\widetilde{k}(x_1, x), \widetilde{k}(x_2, x), \ldots, \widetilde{k}(x_N, x))^T. \end{aligned} \tag{24}$$

3.1. Measure of neighbor samples. In this subsection, we shall introduce six familiar different measures. Let $X = [X_1, X_2, \ldots, X_p]^T$ be a total sample with p features. $\{x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T\}_{i=1}^n$ ($i = 1, 2, \ldots, n$) contains n samples. Every sample can be seen as a point in a p-dimensional space. Let $d(x_i, x_j)$ be a measure between x_i and x_j, where x_i and x_j are samples. Six different measures are used in our paper. In addition, Cover (1965) defined the relation of the sample size and the feature space dimension. The measures in question are

I. Euclidean measure,

$$d(x_i, x_j) = \Big[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\Big]^{1/2},$$

II. absolute measure,

$$d(x_i, x_j) = \sum_{k=1}^{p}|x_{ik} - x_{jk}|,$$

III. Minkowski measure,

$$d(x_i, x_j) = [\sum_{k=1}^{p}|x_{ik} - x_{jk}|^m]^{1/m},$$

where $m \geq 1$,

IV. Chebyshev measure,

$$d(x_i, x_j) = \max_{1\leq k\leq p}|x_{ik} - x_{jk}|,$$

V. minimum measure,

$$d(x_i, x_j) = \min_{1\leq k\leq p}|x_{ik} - x_{jk}|,$$

VI. variance weighted measure,

$$d(x_i, x_j) = \Big[\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{S_k^2}\Big]^{1/2},$$

where

$$S_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2,$$

$$\overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$$

$(i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots, p)$

3.2. Fuzzy k-nearest neighbor algorithm. In our method, fuzzy membership degrees and each class center are obtained with the FKNN algorithm. In addition, there are other similar k-Nearest Neighbor (KNN) methods (Aydilek and Arslan, 2012). With the FKNN algorithm, the computations of the membership degree can be realized through a sequence of steps:

Step 1: Compute six different measure matrices between pairs of feature vectors in the training set.

Step 2: Set the diagonal elements of the six different measure matrices as infinity.

Step 3: Sort the distance matrix (treat each of its columns separately) in ascending order. Collect the class labels of the patterns located in the closest neighborhood of the pattern under consideration (as we are concerned with k neighbors, this returns a list of k integers).

Step 4: Compute the membership degree to Class i for the j-th pattern using the expression proposed by Keller *et al*. (1985),

$$u_{ij} = \begin{cases} 0.51 + 0.49 \times (n_{ij}/k), \text{ if } i \text{ is the same as the } j\text{-th} \\ \text{label of the pattern.} \\ 0.49 \times (n_{ij}/k), \text{ otherwise.} \end{cases}$$

In the above expression n_{ij} stands for the number of the neighbors of the j-th datum (pattern) that belong to the i-th class. As usual, u_{ij} satisfies two obvious properties:

$$\sum_{i=1}^{c} u_{ij} = 1, \quad 0 < \sum_{j=1}^{N} u_{ij} < N.$$

Therefore, the fuzzy membership matrix U can be achieved with the help of the FKNN: $U = [u_{ij}]$ $(i = 1, 2, \ldots, c; j = 1, 2, \ldots, N)$.

3.3. KFDA/QR algorithm. To solve the problem (21), we considered two stages: the first stage is to maximize the pseudo between-class scatter matrix S_b' by the 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:

$$\hat{A} = \arg \max_{\hat{A}^T\hat{A}=I} \operatorname{tr}(\hat{A}^T S_b' \hat{A}). \tag{25}$$

We can see that M is a $c \times c$ block diagonal symmetric matrix. It is easy to decompose M into the form $M = M_1 M_1^T$, where $M_1 = \operatorname{diag}(\sqrt{n_1}, \ldots, \sqrt{n_j}, \ldots, \sqrt{n_c})$ is a $c \times c$ matrix and n_j is the number of training samples in the j-th class. Consequently, $S_b' = (\widetilde{K}U^T M_1)(\widetilde{K}U^T M_1)^T = K_1(K_1)^T$, where K_1 is an $N \times c$ matrix.

In general, the number of classes is smaller than that of training samples. In this case, we can easily prove that $\operatorname{rank}(S_b') \le c - 1$. When c is much smaller than the number of training samples, we can apply the QR technique to decompose K_1 and obtain an efficient method for kernel fuzzy discriminant analysis. In fact, if $K_1 = (Q_1\ Q_2)\binom{R}{0}$ is the QR decomposition of K_1, where $R \in \mathbb{R}^{r\times c}$ is a row full rank matrix, $r = \operatorname{rank}(S_b')$ and $Q_1 \in \mathbb{R}^{N\times r}$ and $Q_2 \in \mathbb{R}^{N\times(N-r)}$ are column orthogonal matrices, we can verify that Q_1 is a solution to the problem (25).

Theorem 1. *For any orthogonal matrix $G \in \mathbb{R}^{r\times r}$, $\hat{A} = Q_1 G$ is a solution to the problem (25).*

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

$$\begin{aligned} \operatorname{tr}((Q_1 G)^T S_b' (Q_1 G)) &= \operatorname{tr}(Q_1^T S_b' Q_1 G G^T) \\ &= \operatorname{tr}(Q_1^T S_b' Q_1), \end{aligned}$$

which indicates that the conclusion is true. ■

Theorem 2. *Let $r = \operatorname{rank}(S_b')$ and $K_1 = Q_1 R$ be the QR decomposition of K_1. Let $\widetilde{S}_t = Q_1^T S_t' Q_1$, $\widetilde{S}_b = Q_1^T S_b' Q_1$ and G be a matrix whose columns are the eigenvectors of $(\widetilde{S}_b)^{-1}\widetilde{S}_t$ corresponding to the t largest eigenvalues. Then $Q_1 G$ is an optimal solution to the problem (21).*

Proof. By the QR decomposition of K_1, we know that $\widetilde{S}_b = Q_1^T S_b' Q_1 = R_1 R_1^T$ is a nonsingular matrix. According to the definition of the pseudo-inverse of a matrix, we can deduce that

$$\begin{aligned} &(S_b')^+ \\ &\quad = (K_1(K_1)^T)^+ \\ &\quad = (\begin{bmatrix} Q_1\ Q_2 \end{bmatrix} \begin{bmatrix} RR^T & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1\ Q_2 \end{bmatrix}^T)^+ \\ &\quad = \begin{bmatrix} Q_1\ Q_2 \end{bmatrix} \begin{bmatrix} (RR^T)^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1\ Q_2 \end{bmatrix}^T. \end{aligned}$$

and then

$$(S_b')^+ S_t' g = (\left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] \left[\begin{array}{cc} (RR^T)^{-1} & 0 \\ 0 & 0 \end{array}\right] \times \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right]^T) S_t' g = \lambda g,$$

which is equivalent to

$$\left[\begin{array}{c} (RR^T)^{-1} \\ 0 \end{array}\right] Q_1^T S_t' \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] \left[\begin{array}{c} Q_1^T \\ Q_2^T \end{array}\right] g = \lambda \left[\begin{array}{c} Q_1^T \\ Q_2^T \end{array}\right] g.$$

Hence,

$$(RR^T)^{-1} Q_1^T S_t' Q_1 Q_1^T g \quad = (\widetilde{S}_b)^{-1} \widetilde{S}_t Q_1^T g = \lambda Q_1^T g,$$

which implies that $Q_1^T g$ is an eigenvector of $(\widetilde{S}_b)^{-1}\widetilde{S}_t$ corresponding to the eigenvalue λ. Therefore, the conclusion of the theorem is true. ■

By Theorem 2, we can propose Algorithm 1.

Algorithm 1. KFDA/QR.

Step 1. Select a measure type from Section 3.1. With the help of the FKNN algorithm, compute the fuzzy membership matrix U.

Step 2. Select a kernel type and compute the kernel matrix K and $\widetilde{K}$.

Step 3. Compute matrices $S_b' = \widetilde{K}U^T M U \widetilde{K}$ and $S_t' = \widetilde{K}\widetilde{K}$.

Step 4. Let $S_b' = K_1 K_1^T$ and calculate the QR decomposition of K_1: $K_1 = Q_1 R$.

Step 5. Let $\widetilde{S}_t = Q_1^T S_t' Q_1$ and $\widetilde{S}_b = Q_1^T S_b' Q_1$.

Step 6. Compute the eigenvectors, denoted by G, of the matrix $(\widetilde{S}_b)^{-1}\widetilde{S}_t$ corresponding to the t largest eigenvalues.

Step 7. Let $A = Q_1 G$.

Step 8. For any input vector x, its low-dimensional feature representation by KFDA/QR is

$$y_x = A^T \widetilde{\Phi}(X)^T \phi(x) = G^T Q_1^T (I - 1_{N\times N}/N)^T (k(x_1, x), \ldots, k(x_N, x))^T.$$

3.4. KFDA/SVD algorithm. To solve the problem (21), we reconsidered the SVD of S_b':

$$S_b' = \left[\begin{array}{cc} U_{b1} & U_{b2} \end{array}\right] \left[\begin{array}{cc} \Sigma_{b1} & 0 \\ 0 & 0 \end{array}\right] \left[\begin{array}{c} U_{b1}^T \\ U_{b2}^T \end{array}\right],$$

where $U_{b1} \in= \mathbb{R}^{N\times r}$ and $U_{b2} \in \mathbb{R}^{N\times(N-r)}$ are column orthogonal matrices, $\Sigma_{b1} \in= \mathbb{R}^{r\times r}$ is a diagonal matrix with non-increasing positive diagonal components and $\mathrm{rank}(S_b') = r$. It is obvious that the matrix $\overline{S}_b = U_{b1}^T S_b' U_{b1} = \Sigma_{b1}$ is nonsingular. Let $\overline{S}_t = U_{b1}^T S_t' U_{b1}$. We thus have Algorithm 2.

Algorithm 2. KFDA/SVD.

Step 1. Select a measure type from Section 3.1. With the help of the FKNN algorithm, compute the fuzzy membership matrix U.

Step 2. Select a kernel type and compute the kernel matrix K and $\widetilde{K}$.

Step 3. Let $S_b' = \widetilde{K}U^T M U \widetilde{K}$ and $S_t' = \widetilde{K}\widetilde{K}$.

Step 4. Compute the SVD of S_b':

$$S_b' = \left[\begin{array}{cc} U_{b1} & U_{b2} \end{array}\right] \left[\begin{array}{cc} \Sigma_{b1} & 0 \\ 0 & 0 \end{array}\right] \left[\begin{array}{c} U_{b1}^T \\ U_{b2}^T \end{array}\right].$$

Step 5. Let $\overline{S}_t = U_{b1}^T S_t' U_{b1}$ and $\overline{S}_b = U_{b1}^T S_b' U_{b1}$.

Step 6. Compute the eigenvectors of the matrix $(\overline{S}_t)^{-1}\overline{S}_b$, denoted by $\widetilde{G}$, corresponding to the t largest eigenvalues.

Step 7. Let $B = U_{b1}\widetilde{G}$.

Step 8. For any input vector x, its low-dimensional feature representation by KFDA/SVD is

$$y_x = B^T \widetilde{\Phi}(X)^T \phi(x) = \widetilde{G}^T U_{b1}^T (I - 1_{N\times N}/N)^T (k(x_1, x), \cdots, k(x_N, x))^T.$$

4. Experiments and analysis

We evaluate the performance of the KFDA/QR and KFDA/SVD algorithms in face recognition tasks. The publicly available face databases, namely, ORL and FERET, are used in the experiments.

All experiments are performed on a PC (2.40 GHZ CPU, 2G RAM) with MATLAB 7.1. Six face recognition methods, namely, KFDA/QR, KFDA/SVD, FDA/QR, FDA/SVD, KDA/QR and KDA/SVD, are tested and compared. For each method, the recognition procedure consists of the following steps:

(i) Six different face recognition procedures are implemented on original images with respect to the number of projection vectors.

(ii) The parameter k of the FKNN algorithm is set to 3, and the parameter m of the Minkowski measure is set to 3.

(iii) The nearest neighbor classifier is used.

It is known that appropriate kernel function selection is important to achieve a better performance

in kernel-based learning methods. Generally speaking, there are two classes of widely used kernel functions: a polynomial kernel and a Gaussian kernel. In order to evaluate the efficiency of QR decomposition and SVD in the KFDA/QR and KFDA/SVD algorithms, respectively, we take into consideration the polynomial kernel (26) and the Gaussian kernel (27). Figure 1 shows a block diagram of our simple system. In addition, the example in Appendix explains how to incorporate the contribution of each training sample into the fuzzy membership matrix U with the help of the measures

$$k(x, y) = (x \cdot y + 1)^p, \tag{26}$$

$$k(x, y) = \exp(-\|x - y\|^2/2\sigma^2). \tag{27}$$

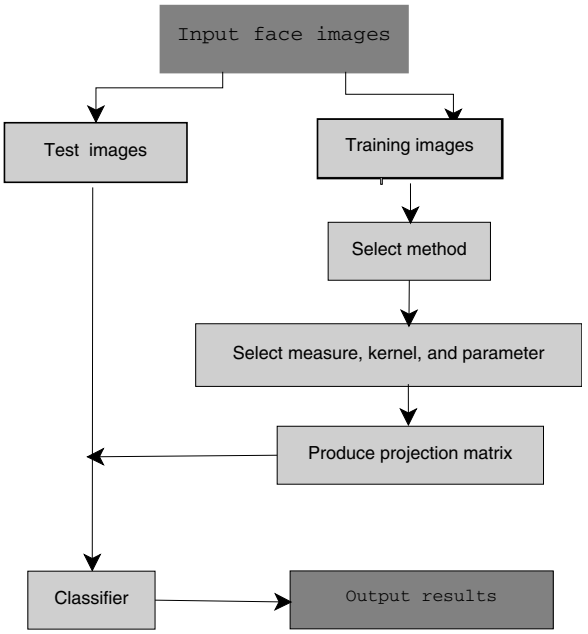


Fig. 1. Simple experiment diagram.

4.1. Experiments with the ORL face database. The ORL face database (Liu, 2006) contains 40 persons, each having 10 different images. 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 the 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. In the experiments, 8 images are randomly taken from 10 images as training samples and the rest are used as testing samples. In order to make full use of the available data and to evaluate the generalization power of the algorithms more accurately, we adopt across-validation strategy and run the system 30 times. Figure 2 shows several sample images of some persons in ORL.

Fig. 2. Sample images of some persons in the ORL database.

In our experiments, the parameters p (from 1 to 6, the step is 0.5) and σ (from 1 to 50, the step is 1) are determined by the across-validation strategy. So, for the ORL database, $p = 2$ and $\sigma = 12$ are the optimal choice.

We tested the performance of KFDA/QR, KFDA/SVD, FDA/QR and FDA/SVD with different measures from Section 3.1. For convenience, the Euclidean measure, absolute measure, Minkowski measure, Chebyshev measure, minimum measure and variance weighted measure are substituted for d1, d2, d3, d4, d5 and d6, respectively. KDA/QR and KDA/SVD do not rely on any measure. This is due to the fact that the contribution of the measure is only made through the fuzzy membership matrix U. The experimental results are shown in Table 1. In addition, in Table 1, the number of projection vectors is 39.

According to Table 1, we have the following conclusion:

(i) For the ORL data set, according to the mean correct recognition rate, KFDA/QR and KFDA/SVD outperform other methods under the Euclidean measure, absolute measure, Chebyshev measure, minimum measure and variance weighted measure with respect to the polynomial kernel $p = 2$.

(ii) According to Standard Deviation (SD), KFDA/QR outperforms KFDA/SVD with respect to six different measures. Therefore, QR decomposition plays an important role in eigenvalue calculation of the matrix.

(iii) For each algorithm, the standard deviation that is obtained by using the QR decomposition method is smaller than that of SVD.

(iv) According to maximum correct recognition rate (MaxR), the minimum measure outperforms other measures. In addition, the highest correct recognition rate of KFDA/QR, KFDA/SVD, FDA/QR, FDA/SVD, KDA/QR and KDA/SVD is

Table 1. Maximum, minimum and mean correct recognition rate (%) (MaxR, MinR and MeanR) of different algorithms and their Standard Deviation (SD) on ORL (polynomial kernel $p = 2$, ME denotes measure).

Algorithm	ME	MaxR	MinR	MeanR (SD)
KFDA/QR		97.21	96.45	**97.08±0.166**
KFDA/SVD	d1	97.21	96.05	97.08±0.263
FDA/QR		97.21	96.05	97.08±0.285
FDA/SVD		96.58	94.89	96.05±0.590
KFDA/QR		97.59	97.01	**97.17±0.161**
KFDA/SVD	d2	97.42	96.41	97.17±0.178
FDA/QR		97.28	96.13	97.12±0.318
FDA/SVD		97.25	94.99	96.58±0.679
KFDA/QR		97.18	96.58	96.92±0.169
KFDA/SVD	**d3**	97.17	96.12	96.92±0.232
FDA/QR		**97.55**	96.46	**97.21±0.256**
FDA/SVD		**97.46**	95.41	96.21±0.442
KFDA/QR		96.80	95.99	**96.42±0.230**
KFDA/SVD	d4	96.86	95.88	96.42±0.277
FDA/QR		96.45	95.59	95.96±0.250
FDA/SVD		96.44	95.47	95.96±0.288
KFDA/QR		**97.70**	96.77	**97.21±0.272**
KFDA/SVD	**d5**	**97.75**	96.78	97.21±0.292
FDA/QR		97.13	96.22	96.75±0.255
FDA/SVD		96.87	96.19	96.46±0.167
KFDA/QR		97.28	97.05	**97.12±0.043**
KFDA/SVD	d6	97.33	96.98	97.12±0.094
FDA/QR		97.47	96.88	97.12±0.140
FDA/SVD		96.78	96.02	96.37±0.238
KDA/QR		97.06	96.12	97.01±0.168
KDA/SVD		78.95	76.11	78.25±0.855

97.70%, 97.75%, 97.55%, 97.46%, 97.06% and 78.95%, respectively.

Further experiments using the Gaussian kernel and different measures will have to be employed to provide a more valuable comparison. The experimental results are given in Table 2.

According to Table 2, we have the following conclusion:

(i) For each method, MeanR will slowly increase as the number of projection vectors increases.

(ii) According to MeanR, for each method which based on QR decomposition, is superior to the ones based on SVD.

(iii) Long projection vectors do not lead to a higher correct recognition rate. Meanwhile, KFDA/QR and KFDA/SVD are sensitive to the measure.

(iv) According to MeanR, for the Gaussian kernel, the minimum measure and variance weighted measure outperform other measures.

Table 2. Mean correct recognition rate (%) of different algorithms on ORL (Gaussian kernel $\sigma = 12$, ME denotes measure, the number of projection vectors is 5, 15, 25 and 35, respectively).

Algorithm	ME	5	15	25	35
KFDA/QR		84.50	96.12	97.67	97.50
KFDA/SVD	d1	85.08	95.42	96.37	97.17
KFDA/QR		84.83	96.17	97.42	97.33
KFDA/SVD	d2	85.29	95.62	96.67	97.29
KFDA/QR		84.25	96.04	97.25	97.58
KFDA/SVD	d3	84.79	95.50	96.42	97.46
KFDA/QR		84.12	94.83	96.79	96.79
KFDA/SVD	d4	83.21	94.04	95.79	96.58
KFDA/QR		83.83	95.58	97.50	97.67
KFDA/SVD	**d5**	85.87	95.00	96.83	**97.71**
KFDA/QR		85.37	96.08	**97.71**	97.42
KFDA/SVD	**d6**	85.21	95.58	96.79	97.21
FDA/QR		85.73	95.94	**97.28**	97.15
FDA/SVD		84.77	95.79	97.00	**97.12**
KDA/QR		82.88	95.75	**97.25**	**97.25**
KDA/SVD		66.92	76.67	77.62	**78.25**

(v) From the MeanR perspective, the correlation between the measure and the correct classification accuracy in FDA is smaller than in KFDA.

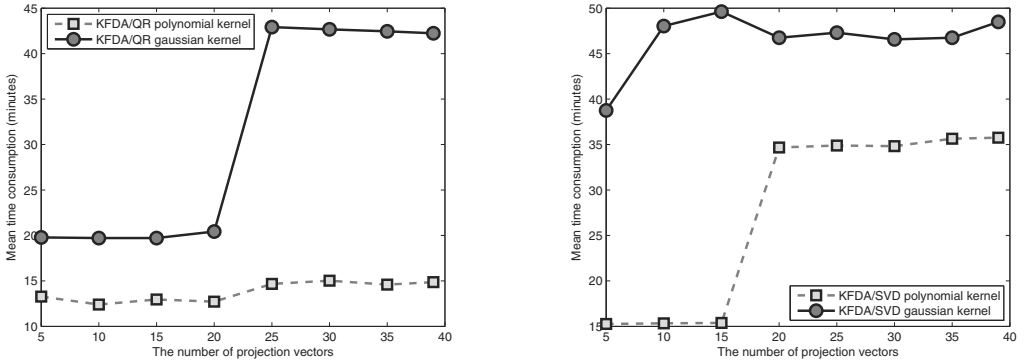
(vi) According to Table 2, the highest correct recognition rates of KFDA/QR, KFDA/SVD, FDA/QR, FDA/SVD, KDA/QR and KDA/SVD are 97.71%, 97.71%, 97.28%, 97.12%, 97.25% and 78.25%, respectively.

In addition, compared with the polynomial kernel, the effectiveness of the Gaussian kernel is significant. However, it takes much more computing time. We perform a contrast test of elapsed times between the polynomial kernel and the Gaussian one. And then, the average time that is obtained by using different measures is recorded. The elapsed times of the polynomial kernel and Gaussian one are listed in Fig. 3.

The results of Fig. 3 clearly show the superiority of the polynomial kernel. In addition, we can see that no matter what kernel (polynomial or Gaussian) we use, the elapsed time of KFDA/QR is less than that of KFDA/SVD. Therefore, in the following experiment, we will explore the performance of different measures with the help of a 2-polynomial kernel (on ORL). The experimental results are shown in Figs. 4–9.

According to the results in Figs. 4–9, the proposed approach outperforms FDA and KDA in terms of the mean correct recognition rate. In addition, we can see that the mean correct recognition rate will slowly increase as the number of projection vectors increases.

MeanR is very affected by what we use as measure on recognition tasks. Meanwhile, we can clearly see that KFDA/QR outperforms FDA/QR and KDA/QR with the

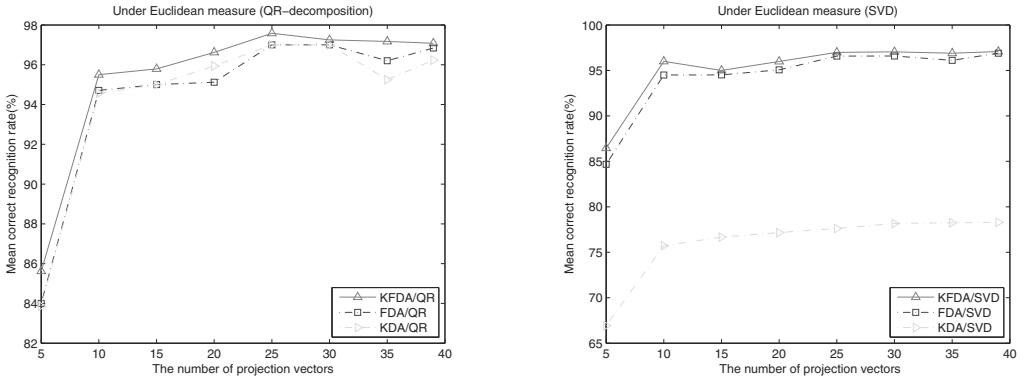


(a) (b)

Fig. 3. Mean time consumption (minutes) on ORL: KFDA/QR (a), KFDA/SVD (b).

Table 3. Different classification methods on ORL using the t-test (reference data $t_{0.05}(29) = 1.699$).

	KFDA/QR vs. FDA/QR	KFDA/QR vs. KDA/QR	KFDA/SVD vs. FDA/SVD	KFDA/SVD vs. KDA/SVD
Null hypothesis	H_0: $u_1 - u_3 \leq 0$	H_0: $u_1 - u_5 \leq 0$	H_0: $u_2 - u_4 \leq 0$	H_0: $u_2 - u_6 \leq 0$
Alternative hypothesis	H_1: $u_1 - u_3 > 0$	H_1: $u_1 - u_5 > 0$	H_1: $u_2 - u_4 > 0$	H_1: $u_2 - u_6 > 0$
d1	$t = 0.05 < 1.699$, **accept H_0, reject H_1**	$t = 1.87 > 1.699$, reject H_0, accept H_1	$t = 9.45 > 1.699$, reject H_0, accept H_1	$t = 114.57 > 1.699$ reject H_0, accept H_1
d2	$t = 0.86 < 1.699$, **accept H_0, reject H_1**	$t = 4.16 > 1.699$, reject H_0, accept H_1	$t = 4.63 > 1.699$, reject H_0, accept H_1	$t = 118.52 > 1.699$ reject H_0, accept H_1
d3	$t = -5.81 < 1.699$, **accept H_0, reject H_1**	$t = -1.97 < 1.699$, **accept H_0, reject H_1**	$t = 7.39 > 1.699$, reject H_0, accept H_1	$t = 115.80 > 1.699$ reject H_0, accept H_1
d4	$t = 7.13 > 1.699$, reject H_0, accept H_1	$t = -11.83 < 1.699$, **accept H_0, reject H_1**	$t = 6.07 > 1.699$, reject H_0, accept H_1	$t = 111.50 > 1.699$ reject H_0, accept H_1
d5	$t = 7.74 > 1.699$, reject H_0, accept H_1	$t = 3.23 > 1.699$, reject H_0, accept H_1	$t = 11.20 > 1.699$, reject H_0, accept H_1	$t = 109.63 > 1.699$ reject H_0, accept H_1
d6	$t = 0.03 < 1.699$, **accept H_0, reject H_1**	$t = 3.77 > 1.699$, reject H_0, accept H_1	$t = 17.44 > 1.699$, reject H_0, accept H_1	$t = 116.27 > 1.699$ reject H_0, accept H_1



(a) (b)

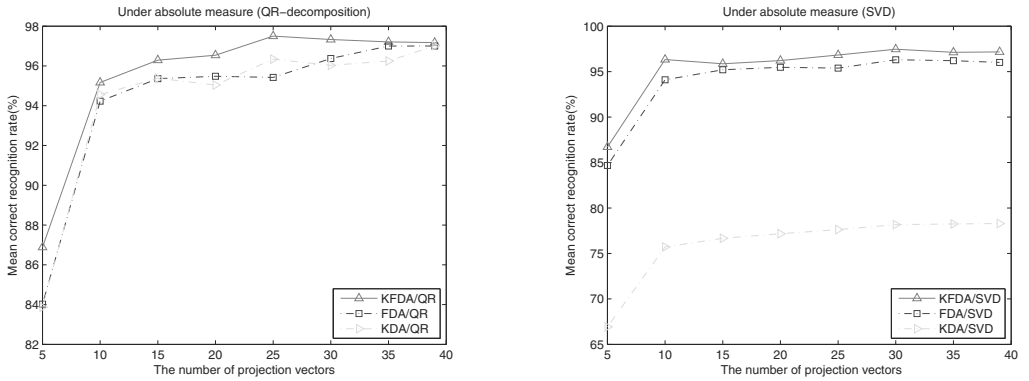
Fig. 4. Mean correct recognition rate curves with the Euclidean measure on ORL: QR decomposition (a), SVD (b).

help of the absolute measure and the variance weighted measure. The main reason is that the fuzzy between-class and within-class scatter matrices are constructed using different fuzzy membership matrices U. Therefore, for recognition tasks, a measure should be considered. It is worth noting that the SVD of KFDA only slightly improves for recognition tasks.

In addition, we found that the differences between the results for different classification methods are very small. So, we need to derive a t-test statistic. Details can be found in the works of Demsar (2006) and Dietterich (1998). The computations of t-test statistics can be realized through a sequence of steps:

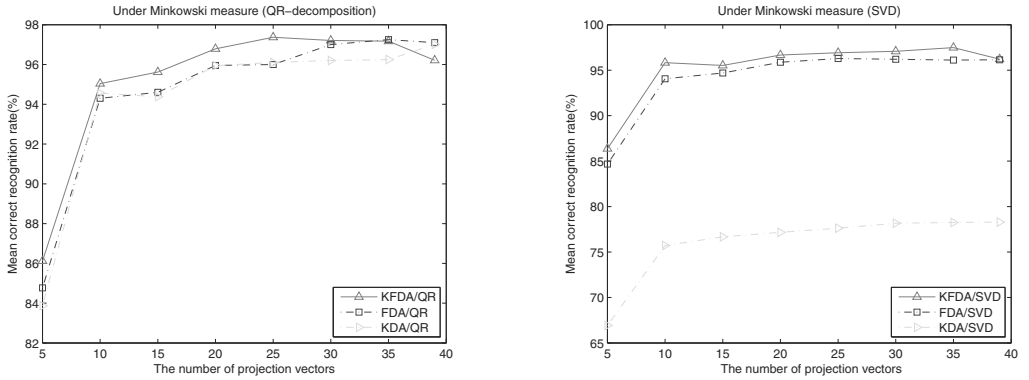
Step 1: Establish a null hypothesis: H_0: $u_1 = u_2$.

Step 2: Compute the t statistic: $t = \overline{d}/(s_d/\sqrt{n})$, where



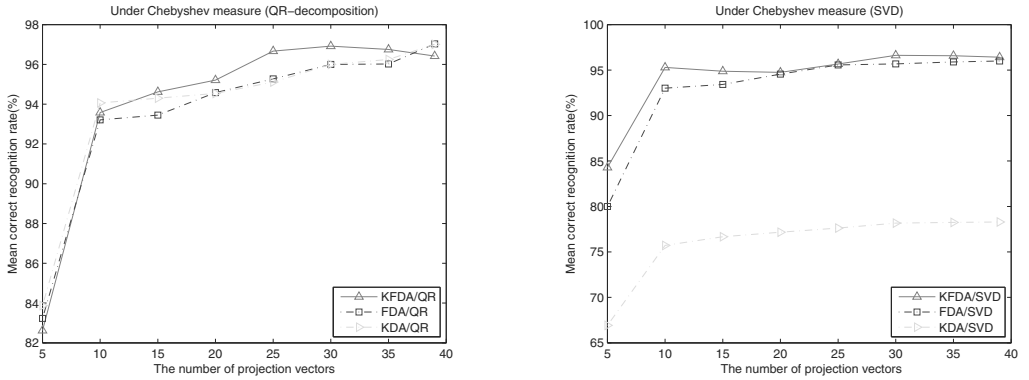
(a) (b)

Fig. 5. Mean correct recognition rate curves with the absolute measure on ORL: QR decomposition (a), SVD (b).



(a) (b)

Fig. 6. Mean correct recognition rate curves with the Minkowski measure on ORL: QR decomposition (a), SVD (b).



(a) (b)

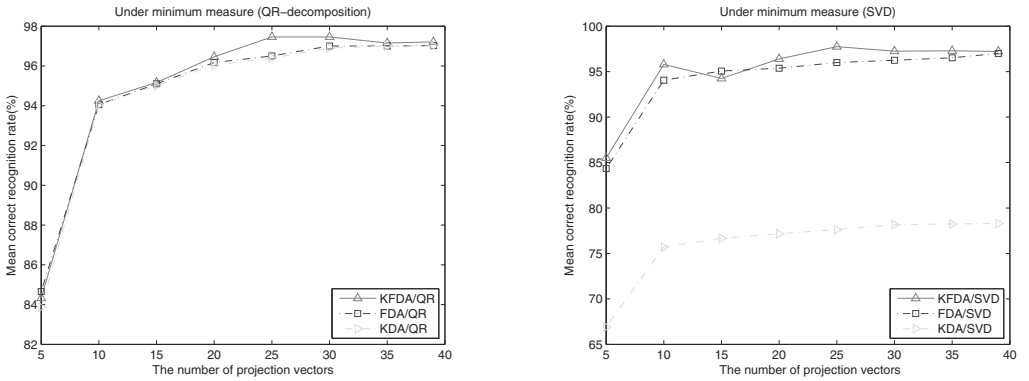
Fig. 7. Mean correct recognition rate curves with the Chebyshev measure on ORL: QR decomposition (a), SVD (b).

$\overline{d}$ is the mean value of differences between the sample, s_d is the standard deviation of difference values, n is the number of samples.

Step 3: According to degrees of freedom, determine the significance level α (generally speaking $\alpha = 0.05$), the look-up table and the contrast.

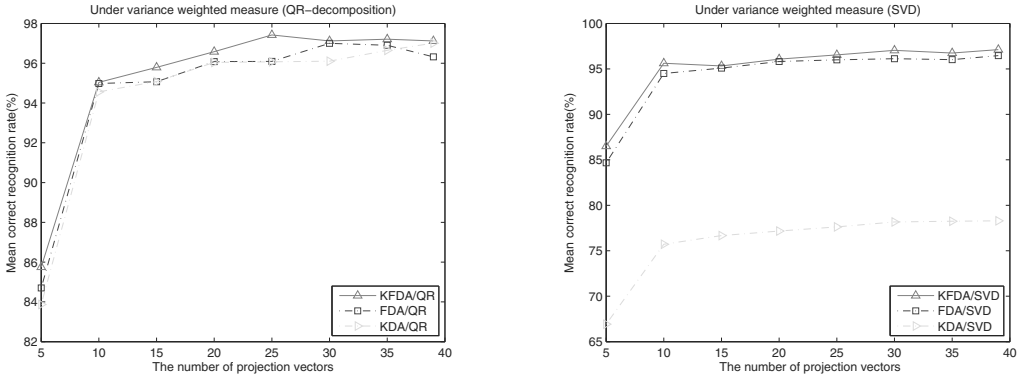
Step 4: According to the t value, we make a decision (reject or accept H_0).

For convenience, the mean recognition rates of KFDA/QR, KFDA/SVD, FDA/QR, FDA/SVD, KDA/QR and KDA/SVD are substituted for u_1, u_2, u_3, u_4, u_5 and u_6, respectively. The results are listed in Table 3. According to the t value of the results in Table 3, we can see that the measure plays an important role in the classification task. Meanwhile, in most cases, the



(a) (b)

Fig. 8. Mean correct recognition rate curves with the minimum measure on ORL: QR decomposition (a), SVD (b).



(a) (b)

Fig. 9. Mean correct recognition rate curves with the variance weighted measure on ORL: QR decomposition (a), SVD (b).

proposed approach outperforms FDA and KDA in terms of the mean correct recognition rate. So, it is necessary to combine the advantages of the fuzzy method and the kernel trick under different measures.

4.2. Experiments with the FERET face database. The FERET face database (Phillips, 2004) is a result of the FERET program, which was sponsored by the US Department of Defense through the DARPA program. The FERET face database contains 14051 gray scale images from 1199 different subjects including pose, facial expression, and illumination variations. In the experiments, 200 individuals with 1000 images are selected. Each person has 5 images. Two images are randomly taken from these images as training samples and the rest are used as testing samples. In order to reduce the size of the image, we obtain the size of 20×20 pixels.

In order to make full use of the available data and to evaluate the generalization power of algorithms more accurately, we adopt a across-validation strategy and run the system thirty times. Figure 10 shows several sample images of some persons in FERET. For the Gaussian kernel, the methods such as KFDA/QR and KFDA/SVD need more elapsed time. Therefore, the polynomial kernel is used for experiments. The parameters p (from 1 to 6, the step is 0.5) are determined by the across-validation strategy. So, for FERET, $p = 6$ is the optimal choice, the number of projection vectors is 199. According to the

Fig. 10. Sample images of some persons in FERET.

results in Table 5, we have the following conclusions:

(i) For the FERET data set, according to MeanR

Table 5. Different classification methods on FERET using the t-test (reference data $t_{0.05}(29) = 1.699$).

	KFDA/QR vs. FDA/QR	KFDA/QR vs. KDA/QR	KFDA/SVD vs. FDA/SVD	KFDA/SVD vs. KDA/SVD
Null hypothesis	H_0: $u_1 - u_3 \leq 0$	H_0: $u_1 - u_5 \leq 0$	H_0: $u_2 - u_4 \leq 0$	H_0: $u_2 - u_6 \leq 0$
Alternative hypothesis	H_1: $u_1 - u_3 > 0$	H_1: $u_1 - u_5 > 0$	H_1: $u_2 - u_4 > 0$	H_1: $u_2 - u_6 > 0$
d1	$t = -45.82 < 1.699$, **accept H_0, reject H_1**	$t = 109.44 > 1.699$, reject H_0, accept H_1	$t = 65.08 > 1.699$, reject H_0, accept H_1	$t = 227.91 > 1.699$ reject H_0, accept H_1
d2	$t = -44.18 < 1.699$, **accept H_0, reject H_1**	$t = 112.39 > 1.699$, reject H_0, accept H_1	$t = 115.38 > 1.699$, reject H_0, accept H_1	$t = 259.73 > 1.699$ reject H_0, accept H_1
d3	$t = -22.52 < 1.699$, **accept H_0, reject H_1**	$t = 105.26 > 1.699$, reject H_0, accept H_1	$t = 66.85 > 1.699$, reject H_0, accept H_1	$t = 230.15 > 1.699$ reject H_0, accept H_1
d4	$t = 62.94 > 1.699$, reject H_0, accept H_1	$t = 70.56 > 1.699$, reject H_0, accept H_1	$t = 78.78 > 1.699$, reject H_0, accept H_1	$t = 212.55 > 1.699$ reject H_0, accept H_1
d5	$t = 253.80 > 1.699$, reject H_0, accept H_1	$t = 72.45 > 1.699$, reject H_0, accept H_1	$t = 47.83 > 1.699$, reject H_0, accept H_1	$t = 209.00 > 1.699$ reject H_0, accept H_1
d6	$t = -51.68 < 1.699$, **accept H_0, reject H_1**	$t = 89.01 > 1.699$, reject H_0, accept H_1	$t = 89.61 > 1.699$, reject H_0, accept H_1	$t = 236.22 > 1.699$ reject H_0, accept H_1

Table 4. Maximum, minimum and mean correct recognition rate (%) (MaxR, MinR and MeanR) of different algorithms and their Standard Deviation (SD) on FERET (polynomial kernel $p = 6$, ME denotes measure).

Algorithm	ME	MaxR	MinR	MeanR (SD)
KFDA/QR		51.25	51.10	51.16±0.444
KFDA/SVD	d1	**51.30**	50.80	51.16±0.177
FDA/QR		**54.22**	53.27	**53.73±0.310**
FDA/SVD		45.85	45.45	45.62±0.122
KFDA/QR		**52.10**	51.86	51.99±0.072
KFDA/SVD	d2	51.29	50.86	51.09±0.131
FDA/QR		53.45	53.03	**53.23±0.123**
FDA/SVD		45.86	45.13	45.46±0.221
KFDA/QR		51.16	51.01	51.09±0.044
KFDA/SVD	d3	51.29	50.88	51.08±0.139
FDA/QR		52.67	51.85	**52.24±0.279**
FDA/SVD		**46.98**	44.09	45.52±0.454
KFDA/QR		49.96	49.78	**49.88±0.062**
KFDA/SVD	**d4**	50.19	49.58	49.88±0.181
FDA/QR		45.97	44.25	45.38±0.403
FDA/SVD		43.88	42.11	42.16±0.422
KFDA/QR		49.64	49.50	**49.57±0.040**
KFDA/SVD	**d5**	49.98	48.74	49.57±0.277
FDA/QR		44.48	44.15	44.31±0.108
FDA/SVD		46.74	46.01	46.32±0.226
KFDA/QR		50.90	50.58	50.74±0.110
KFDA/SVD	d6	51.08	50.38	50.70±0.210
FDA/QR		52.47	52.02	**52.29±0.131**
FDA/SVD		45.72	45.01	45.28±0.205
KDA/QR		46.74	45.94	46.33±0.249
KDA/SVD		32.64	30.21	31.23±0.422

and MaxR, FDA/QR outperforms other methods. However, FDA/SVD does not have a strong advantage for recognition problems. The reason may be in the low efficiency of projection vectors that is obtained by SVD.

(ii) KFDA outperforms other methods with the help of the Chebyshev measure and minimum measure.

(iii) For each algorithm, SD that is obtained by using the QR decomposition method is smaller than that of SVD (except for the Euclidean measure). In addition, KFDA/QR and KFDA/SVD are more sensitive to the measure.

(iv) The highest correct recognition rate of KFDA/QR, KFDA/SVD, FDA/QR, FDA/SVD, KDA/QR and KDA/SVD is 52.10%, 51.30%, 54.22%, 46.98%, 46.74% and 32.64%, respectively.

(v) For the FERET data set, the MaxR of every method is not high, mainly because of the following reasons: on the one hand, a small amount of training samples are used, while on the other, long projection vectors are used. For the FERET data set, we also derive a t-test statistic, as shown in Table 4. We can see that the t value change is big. This suggests that there are very big differences between the mean correct recognition rates of each method.

5. Conclusion and future work

In this paper, we proposed two kinds of methods of kernel-based fuzzy discriminant analysis: KFDA/QR and KFDA/SVD for feature extraction with combination of different measures via QR decomposition and the singular value decomposition technique. Through the two methods we can find lower-dimensional nonlinear features with significant discriminant power, and the two methods can be viewed as a generalization of FDA and KDA. In the proposed method, the fuzzy membership degree matrix U that is obtained by combining the measure of features of samples data is incorporated into the definition of between-class and within-class scatter matrices to get fuzzy between-class and within-class scatter matrices. So, for us, how to incorporate the contribution of each training sample into the fuzzy membership matrix U with the help of the measure is a research priority. Experimental results confirm that KFDA is not only feasible, but also achieves

a better recognition performance in the ORL and FERET face databases in terms of the mean correct recognition rate.

Therefore, in order to improve the recognition rate, we should consider the effect of the measure. The future work on this subject will still be investigation of the influence of the measure and the kernel parameter on classification and recognition tasks. In addition, exploring new algorithms to solve the corresponding optimization problems is also a further research direction.

Acknowledgment

First of all, we thank the anonymous reviewers for their constructive comments and suggestions. This work is supported by the National Natural Science Foundation of the PR China (10871226), the Natural Science Foundation of the Shandong Province (ZR2009AL006) and the Young and Middle-Aged Scientists Research Foundation of the Shandong Province (BS2010SF004), PR China, as well as the Graduates' Research Innovation Program of Higher Education of the Jiangsu Province (CXZZ13-0239, CXZZ13-0261, CXZZ13-0932), PR China.

References

Aydilek, I.B. and Arslan, A. (2012). A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks, *International Journal of Innovative Computing, Information and Control* **7**(8): 4705–4717.

Baudat, G. and Anouar, F. (2000). Generalized discriminant analysis using a kernel approach, *Neural Computation* **12**(10): 2385–2404.

Belhumeur, P.N. and Kriegman, D.J. (1997). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis Machine Intelligence* **19**(7): 711–720.

Chen, L.F., Liao, H. Y.M., Ko, M.T., Lin, J.C. and Yu, G.J. (2000). A new LDA-based face recognition system which can solve the small sample size problem, *Pattern Recognition* **33**(10): 1713–1726.

Cover, T.M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Transactions on Electronic Computers* **14**(3): 326–334.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets, *The Journal of Machine Learning Research* **7**(1): 1–30.

Dietterich, T.G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Computation* **10**(7): 1895–1923.

Duda, R.O., Hart, P.E. and Stork, D.G. (2012). *Pattern Classification*, John Wiley and Sons, New York, NY.

Friedman, J.H. (1989). Regularized discriminant analysis, *Journal of the American Statistical Association* **84**(405): 165–175.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, San Diego, California, CA.

Gao, J. and Fan, L. (2011). Kernel-based weighted discriminant analysis with QR decomposition and its application face recognition, *WSEAS Transactions on Mathematics* **10**(10): 358–367.

Gao, J., Fan, L. and Xu, L. (2012). Solving the face recognition problem using QR factorization, *WSEAS Transactions on Mathematics* **11**(1): 728–737.

Gao, J.Q., Fan, L.Y. and Xu, L.Z. (2013). Median null (sw)-based method for face feature recognition, *Applied Mathematics and Computation* **219**(12): 6410–6419.

Gao, Q.X., Zhang, L. and Zhang, D. (2008). Face recognition using FLDA with single training image per person, *Applied Mathematics and Computation* **205**(2): 726–734.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *The Annals of Statistics* **23**(1): 73–102.

Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring, *Journal of the American Statistical Association* **89**(428): 1255–1270.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (1991). The elements of statistical learning: Data mining, inference and prediction, *The Mathematical Intelligencer* **27**(2): 83–85.

Hong, Z.Q. and Yang, J.Y. (2005). Optimal discriminant plane for a small number of samples and design method of classifier on the plane, *Pattern Recognition* **24**(4): 317–324.

Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(2): 153–158.

Keller, J.M., Gray, M.R. and Givens, J.A. (1985). A fuzzy k-nearest neighbor algorithm, *IEEE Transactions on Systems, Man and Cybernetics* **15**(4): 580–585.

Koc, M. and Barkana, A. (2011). A new solution to one sample problem in face recognition using FLDA, *Applied Mathematics and Computation* **217**(24): 10368–10376.

Kwak, K.C. and Pedrycz, W. (2005). Face recognition using a fuzzy Fisherface classifier, *Pattern Recognition* **38**(10): 1717–1732.

Lee, H.M., Chen, C.M., Chen, J.M. and Jou, Y.L. (2001). An efficient fuzzy classifier with feature selection based on fuzzy entropy, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **31**(3): 426–432.

Liu, F. and Xue, X. (2012). Constructing kernels by fuzzy rules for support vector regressions, *International Journal of Innovative Computing, Information and Control* **8**(7): 4811–4822.

Liu, Y. (2006). Website of the ORL face database, `http://www.cam-orl.co.uk`.

Liu, Y., Liu, X. and Su, Z. (2008). A new fuzzy approach for handling class labels in canonical correlation analysis, *Neurocomputing* **71**(7): 1735–1740.

Loog, M., Duin, R.P.W. and Haeb-Umbach, R. (2001). Multiclass linear dimension reduction by weighted pairwise fisher criteria, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(7): 762–766.

Pahasa, J. and Ngamroo, I. (2012). PSO based kernel principal component analysis and multi-class support vector machine for power quality problem classification, *International Journal of Innovative Computing, Information and Control* **8**(3): 1523–1539.

Pal, N.R. and Eluri, V.K. (1998). Two efficient connectionist schemes for structure preserving dimensionality reduction, *IEEE Transactions on Neural Networks* **9**(6): 1142–1154.

Phillips, P.J. (2004). Website of the facial recognition technology (FERET) database, `http://www.itl.nist.gov/iad/humanid/feret/feret-master.html`.

Raudys, S.J. and Jain, A.K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(3): 252–264.

Schölkopf, B., Smola, A. and Müller, K.R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* **10**(5): 1299–1319.

Swets, D.L. and Weng, J.J. (1996). Using discriminant eigenfeatures for image retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(8): 831–836.

Świercz, E. (2010). Classification in the Gabor time-frequency domain of non-stationary signals embedded in heavy noise with unknown statistical distribution, *International Journal of Applied Mathematics and Computer Science* **20**(1): 135–147, DOI: 10.2478/v10006-010-0010-x.

Vapnik, V.N. (1998). *Statistical Learning Theory*, Wiley, New York, NY.

Woźniak, M. and Krawczyk, B. (2012). Combined classifier based on feature space partitioning, *International Journal of Applied Mathematics and Computer Science* **22**(4): 855–866, DOI: 10.2478/v10006-012-0063-0.

Wu, X.H. and Zhou, J.J. (2006). Fuzzy discriminant analysis with kernel methods, *Pattern Recognition* **39**(11): 2236–2239.

Yang, J., Frangi, A.F., Yang, J.Y., Zhang, D. and Jin, Z. (2005). KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(2): 230–244.

Yang, J. and Yang, J.Y. (2001). An optimal FLD algorithm for facial feature extraction, *Proceedings of SPIE Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision, Boston, MA, USA,* pp. 438-444.

Yang, J. and Yang, J.Y. (2003). Why can LDA be performed in PCA transformed space?, *Pattern Recognition* **36**(2): 563–566.

Yang, W., Wang, J., Ren, M., Zhang, L. and Yang, J. (2009). Feature extraction using fuzzy inverse FDA, *Neurocomputing* **73**(13): 3384–3390.

Yu, H. and Yang, J. (2001). A direct LDA algorithm for high-dimensional data with application to face recognition, *Pattern Recognition* **34**(10): 2067–2070.

Zadeh, L.A. (1965). Fuzzy sets, *Information and Control* **8**(3): 338–353.

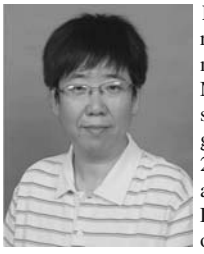
Zheng, Y.J., Yang, J., Yang, J.Y. and Wu, X.J. (2006a). A reformative kernel Fisher discriminant algorithm and its application to face recognition, *Neurocomputing* **69**(13): 1806–1810.

Zheng, Y., Yang, J., Wang, W., Wang, Q., Yang, J. and Wu, X. (2006b). Fuzzy kernel Fisher discriminant algorithm with application to face recognition, *6th World Congress on Intelligent Control and Automation WCICA, Dalian, China*, Vol. 2, pp. 9669–9672.

Zhuang, X.S. and Dai, D.Q. (2005). Inverse Fisher discriminate criteria for small sample size problem and its application to face recognition, *Pattern Recognition* **38**(11): 2192–2194.

Zhuang, X.S. and Dai, D.Q. (2007). Improved discriminate analysis for high-dimensional data and its application to face recognition, *Pattern Recognition* **40**(5): 1570–1578.

Jian-qiang Gao was born in Shandong, China, in 1982. He received his B.Sc. and M.Sc. degrees from Liaocheng University, Shandong, in 2009 and 2012, respectively. Now he is studying for his Ph.D. degree in remote sensor and remote control, remote sensing image classification, information processing system and pattern recognition at the School of College of Computer and Information Engineering, Hohai University, Nanjing, China.



Li-ya Fan was born in 1963 in China. She received her B.Sc. degree from Northeast Normal University, Changchun, China, in 1984, the M.Sc. degree from Inner Mongolia University, Hohhot, China, in 2000, and the Ph.D. degree from Xidian University, Xi'an, China, in 2003. Now, she is a professor of the School at Mathematical Sciences, Liaocheng University, Liaocheng, China. Her research interests include optimization theory and applications, machine learning theory and pattern recognition.

Li Li was born in Shandong, China, in 1985. She received her B.Sc. degree from Liaocheng University, Shandong, in 2012. Now she is studying for her M.Sc. degree in nonlinear analysis and economic application at the School of Department of Mathematics, Nanjing University of Finance and Economics, Nanjing, China.

Li-zhong Xu is a professor in the College of Computer and Information Engineering, as well as the director of the Engineering Research Center of Sensing and Computing, Hohai University, Nanjing, China. He received his Ph.D. degree from the China University of Mining and Technology, Xuzhou, China, in 1997. He is a senior member of the Chinese Institute of Electronics and the China Computer Federation. His current research areas include multi-sensor system and information fusion, signal processing in remote sensing and remote control, information processing system and its applications, system modeling and simulation.

Appendix

Here we give a simple example which explains how to incorporate the contribution of each training sample into the fuzzy membership matrix U with the help of a measure. The Chebyshev and minimum measures are used for test. We consider nine two-dimensional patterns belonging to three-classes as below:

$S =$

No.	Feature1	Feature2	Class
1	0.3	0.4	1
2	0.4	0.2	1
3	0.3	0.2	1
4	0.5	0.4	2
5	0.6	0.5	2
6	0.6	0.4	2
7	0.9	0.7	3
8	0.8	0.7	3
9	0.9	0.8	3

The distance matrix comes with the following entries (employing the Chebyshev measure):

$A1 =$

No.	1	2	3	4	5	6	7	8	9
1	0	0.2	0.2	0.2	0.3	0.3	0.6	0.5	0.6
2	0.2	0	0.1	0.2	0.3	0.2	0.5	0.5	0.6
3	0.2	0.1	0	0.2	0.3	0.3	0.6	0.5	0.6
4	0.2	0.2	0.2	0	0.1	0.1	0.4	0.3	0.4
5	0.3	0.3	0.3	0.1	0	0.1	0.3	0.2	0.3
6	0.3	0.2	0.3	0.1	0.1	0	0.3	0.3	0.4
7	0.6	0.5	0.6	0.4	0.3	0.3	0	0.1	0.1
8	0.5	0.5	0.5	0.3	0.2	0.3	0.1	0	0.1
9	0.6	0.6	0.6	0.4	0.3	0.4	0.1	0.1	0

The diagonal elements are replaced by infinity (Inf):

$A2 =$

No.	1	2	3	4	5	6	7	8	9
1	Inf	0.2	0.2	0.2	0.3	0.3	0.6	0.5	0.6
2	0.2	Inf	0.1	0.2	0.3	0.2	0.5	0.5	0.6
3	0.2	0.1	Inf	0.2	0.3	0.3	0.6	0.5	0.6
4	0.2	0.2	0.2	Inf	0.1	0.1	0.4	0.3	0.4
5	0.3	0.3	0.3	0.1	Inf	0.1	0.3	0.2	0.3
6	0.3	0.2	0.3	0.1	0.1	Inf	0.3	0.3	0.4
7	0.6	0.5	0.6	0.4	0.3	0.3	Inf	0.1	0.1
8	0.5	0.5	0.5	0.3	0.2	0.3	0.1	Inf	0.1
9	0.6	0.6	0.6	0.4	0.3	0.4	0.1	0.1	Inf

Next, the distance matrix is sorted (which is done separately for each column of the matrix):

$A3 =$

No.	1	2	3	4	5	6	7	8	9
1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
3	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.3
4	0.3	0.2	0.3	0.2	0.3	0.3	0.3	0.3	0.4
5	0.3	0.3	0.3	0.2	0.3	0.3	0.4	0.3	0.4
6	0.5	0.5	0.5	0.3	0.3	0.3	0.5	0.5	0.6
7	0.6	0.5	0.6	0.4	0.3	0.3	0.6	0.5	0.6
8	0.6	0.6	0.6	0.4	0.3	0.4	0.6	0.5	0.6
9	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

If we consider $k = 3$ neighbors, then the classes of the i-th nearest point of the j-th input vector are as follows:

$A4 =$

Class	1	2	3	4	5	6	7	8	9
1	2	2	2	1	0	1	0	0	0
2	1	1	1	2	2	2	1	1	1
3	0	0	0	0	1	0	2	2	2

Similarly to the above procedure, the distance matrix comes with the following entries (employing the minimum measure):

$B1 =$

No.	1	2	3	4	5	6	7	8	9
1	0	0.1	0	0	0.1	0	0.3	0.3	0.4
2	0.1	0	0	0.1	0.2	0.2	0.5	0.4	0.5
3	0	0	0	0.2	0.3	0.2	0.5	0.5	0.6
4	0	0.1	0.2	0	0.1	0	0.3	0.3	0.4
5	0.1	0.2	0.3	0.1	0	0	0.2	0.2	0.3
6	0	0.2	0.2	0	0	0	0.3	0.2	0.3
7	0.3	0.5	0.5	0.3	0.2	0.3	0	0	0
8	0.3	0.4	0.5	0.3	0.2	0.2	0	0	0.1
9	0.4	0.5	0.6	0.4	0.3	0.3	0	0.1	0

The diagonal elements are replaced by infinity (Inf):

$B2 =$

No.	1	2	3	4	5	6	7	8	9
1	Inf	0.1	0	0	0.1	0	0.3	0.3	0.4
2	0.1	Inf	0	0.1	0.2	0.2	0.5	0.4	0.5
3	0	0	Inf	0.2	0.3	0.2	0.5	0.5	0.6
4	0	0.1	0.2	Inf	0.1	0	0.3	0.3	0.4
5	0.1	0.2	0.3	0.1	Inf	0	0.2	0.2	0.3
6	0	0.2	0.2	0	0	Inf	0.3	0.2	0.3
7	0.3	0.5	0.5	0.3	0.2	0.3	Inf	0	0
8	0.3	0.4	0.5	0.3	0.2	0.2	0	Inf	0.1
9	0.4	0.5	0.6	0.4	0.3	0.3	0	0.1	Inf

Next, the distance matrix is sorted (which is done separately for each column of the matrix):

$B3 =$

No.	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0.1	0	0	0.1	0	0	0.1	0.1
3	0	0.1	0.2	0.1	0.1	0	0.2	0.2	0.3
4	0.1	0.2	0.2	0.1	0.2	0.2	0.3	0.2	0.3
5	0.1	0.2	0.3	0.2	0.2	0.2	0.3	0.3	0.4
6	0.3	0.4	0.5	0.3	0.2	0.2	0.3	0.3	0.4
7	0.3	0.5	0.5	0.3	0.3	0.3	0.5	0.4	0.5
8	0.4	0.5	0.6	0.4	0.3	0.3	0.5	0.5	0.6
9	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

If we consider $k = 3$ neighbors, then the classes of the i-th nearest point of the j-th input vector are as follows:

$B4 =$

Class	1	2	3	4	5	6	7	8	9
1	1	2	2	2	1	1	0	0	0
2	2	1	1	1	2	2	1	1	1
3	0	0	0	0	0	0	2	2	2

According to the results, we can see that A4 is totally different from B4. Meanwhile, A4 and B4 constitute the fuzzy membership matrix U. Therefore, different measures will affect the composition of between-class and within-class scatter matrices.

Received: 14 March 2013
Revised: 10 July 2013
Re-revised: 5 September 2013