International Journal of Innovative
Computing, Information and Control
Volume **9**, Number **1**, January **2013**

ICIC International ©2013 ISSN 1349-4198
pp. 437–450

ADAPTIVE QUASICONFORMAL KERNEL FISHER DISCRIMINANT ANALYSIS VIA WEIGHTED MAXIMUM MARGIN CRITERION

Chuang Lin[1], Binghui Wang[1], Zheming Lu[2,*] and Kuanjiu Zhou[1]

[1]School of Software
Dalian University of Technology
8# Road, Economy and Technology Development Area, Dalian 116620, P. R. China
{ linchuang_78; zhoukj }@dlut.edu.cn; binghui.wang89@gmail.com

[2]School of Aeronautics and Astronautics
Zhejiang University
No. 38, Zheda Road, Hangzhou 310027, P. R. China
*Corresponding author: zheminglu@zju.edu.cn

Received November 2011; revised March 2012

Abstract. *Kernel Fisher discriminant analysis (KFD) is an effective method to extract nonlinear discriminant features of input data using the kernel trick. However, conventional KFD algorithms endure the kernel selection problem as well as the singular problem. In order to overcome these limitations, a novel nonlinear feature extraction method called adaptive quasiconformal kernel Fisher discriminant analysis (AQKFD) via weighted maximum margin criterion (WMMC) is proposed in this paper. AQKFD, which solves the kernel selection problem, maps the data from the original input space into the quasiconformal kernel mapping space using a quasiconformal kernel. The adaptive parameters of the quasiconformal kernel are calculated through maximizing the measure of class separability of the input data in the quasiconformal kernel mapping space via WMMC which is in terms of the Fisher discriminant criterion. Moreover, when the weight parameter is approximate to the maximum value of Fisher discriminant criterion, then nonlinear features extracted by AQKFD-WMMC have the optimal class separability and AQKFD-WMMC can also solve the singular problem which is endured by KFD. Experimental results on the three real-world datasets, i.e., ORL, YALE and FERET face databases demonstrate the effectiveness of the proposed method.*

Keywords: Quasiconformal, Adaptive quasiconformal kernel, Weighted maximum margin criterion, Kernel Fisher discriminant analysis, Class separability, Feature extraction

1. **Introduction.** Over the last few years, kernel learning or kernel machine has aroused broad interest in pattern recognition and machine learning areas [2]. KPCA was originally developed by Scholkopf et al. in 1998 [3], while KFD was first proposed by Mika et al. in 1999 [4,5]. KFD has been found to be very effective in many real-world applications owing to its good performance on feature extraction. Researchers have developed a series of KFD algorithms (Lu [6], Baudat and Anouar [7], Liang et al. [8-10], Yang et al. [11,12], Zheng et al. [13], Huang et al. [14], Wang et al. [15], Ma et al. [16], Chen et al. [17], Liang et al. [18], Tao et al. [19], Xu et al. [20], Yeung et al. [21], Saadi et al. [22], Liu et al. [23], Shen et al. [24], Wu et al. [25]). For classification problem based on supervised kernel learning, different kernel geometrical structures give different class discrimination. However, the separability of the data in the feature space could be even worse if an inappropriate kernel is used since the geometrical structure of the mapped data in the feature space is totally determined by the kernel matrix, so the selection of kernel influences greatly the performance of KFD. Recently, many methods of optimizing

the kernel parameters of the kernel function are developed. Cristianini et al. [26] and Lanckriet et al. [27] have proposed methods of choosing kernel by optimizing the measure of data separation in the feature space for the first time. The authors respectively use the alignment and employ the margin as the measure of data separation to evaluate the adaptability of a kernel to input data. Huang et al. [14] and Wang et al. [15] respectively use kernel subspace LDA and biased KFD to optimize kernel parameters. Kim et al. [33] illustrate that the optimal kernel selection problem can be reformulated as a tractable convex optimization problem. Hamsici et al. [34] define sparse kernel to optimize Bayes Discriminant Analysis. You et al. [35] derive the first criterion that specifically aims to find a kernel representation where the Bayes classifier becomes linear. However, considering that optimizing kernel parameters just from a set of discrete values of the parameters cannot change the geometrical structures of the data in the kernel mapping space [14,15]. We introduce a so-called adaptive quasiconformal kernel which was studied in the previous work [27,30], where the geometrical structure of data in the feature space is changeable with the different parameters of the quasiconformal kernel. The optimal parameters are calculated through maximizing the class separability of the data in the kernel mapping space via weighted maximum margin criterion which is in terms of the Fisher discriminant criterion, and thus, AQKFD-WMMC is more adaptive to the input data for classification than KFD and also avoids the troublesome singular problem. Experiments on ORL, Yale and FERET face databases demonstrate the superiority of the proposed algorithm compared with conventional KFD algorithm.

The rest of the paper is organized as follows. Section 2 and Section 3 describe the detailed information of AQKFD and WMMC respectively. Experimental results of the proposed algorithm conducted on three real-world datasets are reported in Section 4. Finally, Section 5 comes to a conclusion and offers a discussion.

2. Adaptive Quasiconformal Kernel Fisher Discriminant Analysis.

In this section, we will report the algorithm of AQKFD. Before performing it, we firstly review the basic kernel Fisher discriminant analysis.

2.1. Kernel Fisher discriminant analysis (KFD).

Suppose there are N training samples $x_1, x_2, \cdots, x_N$ with C known pattern classes in n-dimension space $\mathbb{R}^n$, each class has N_i $(i = 1, 2, \ldots, C)$ training samples. For a given nonlinear mapping Φ, the input data space $\mathbb{R}^n$ can be mapped into the feature space $\mathbb{R}^m$:

$$\Phi : \mathbb{R}^n \to \mathbb{R}^m, \quad x \to \Phi(x) \tag{1}$$

As a result, a pattern in the original input data space $\mathbb{R}^n$ is mapped into a potentially much higher dimensional feature vector in the feature space $\mathbb{R}^m$ and Fisher criterion in the feature space $\mathbb{R}^m$ is defined by

$$J_F^{\Phi}(\omega) = \frac{\omega^T S_b^{\Phi} \omega}{\omega^T S_w^{\Phi} \omega} \tag{2}$$

where ω is the vector, $S_b^{\Phi} = \sum_{i=1}^{C} N_i (m_i^{\Phi} - m_0^{\Phi})(m_i^{\Phi} - m_0^{\Phi})^T$, $S_w^{\Phi} = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (\Phi(x_j^i) - m_i^{\Phi})(\Phi(x_j^i) - m_i^{\Phi})^T$ are the between-class scatter matrix and within-class scatter matrix in the feature space, respectively. $\Phi(x_j^i)$ denotes the j-th mapped training sample in class i, $m_i^{\Phi} = 1/N_i \sum_{k=1}^{N_i} \Phi(x_k)$, $m_0^{\Phi} = 1/N \sum_{j=1}^{N} \Phi(x_j)$ are the mean of the mapped training samples in class i and mean across all mapped vector training samples.

The solution of Fisher criterion is to find discriminant vector ω which satisfies

$$\omega^* = \arg\max_{\omega} J_F^{\Phi}(\omega) \tag{3}$$

Theorem 2.1. *The Fisher criterion (FC) function t in Equation (2) can be replaced by*

$$\tilde{J}_F^{\Phi}(\varpi) = \frac{\varpi^T S_b^{\Phi} \varpi}{\varpi^T S_w^{\Phi} \varpi + \varpi^T S_b^{\Phi} \varpi} = \frac{\varpi^T S_b^{\Phi} \varpi}{\varpi^T S_t^{\Phi} \varpi} \tag{4}$$

in the course of solving the discriminant vectors of the optimal set.

From Theorem 2.1, we know that $J_F^{\Phi}(\omega)$ and $\tilde{J}_F^{\Phi}(\varpi)$ are functionally equivalent in terms of solving the optimal set of discriminant vectors. In this paper, we calculate the optimal discriminant vectors ϖ^* based on $\tilde{J}_F^{\Phi}(\varpi)$.

According to the Mercer kernel function theory, any solution ϖ^* of $\tilde{J}_F^{\Phi}(\varpi)$ can be linearly expanded by all mapped training samples in the feature space, that is

$$\varpi = \sum_{p=1}^{N} \alpha_p \Phi(x_p) = \Phi(X)\boldsymbol{\alpha} \tag{5}$$

where $\Phi(X) = [\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N)]$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$. Suppose the data are centered, that is, $m_0^{\Phi} = \mathbf{0}$, then $S_b^{\Phi} = \Phi(X)W(\Phi(X))^T$, $S_t^{\Phi} = \Phi(X)(\Phi(X))^T$, and FC in (4) can be transformed into

$$\tilde{J}_F^{\Phi}(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T KWK\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T KK\boldsymbol{\alpha}} \tag{6}$$

where $W = diag(W_1, W_2, \ldots, W_C)$, and W_i is an $N_i \times N_i$ matrix whose elements are $1/N_i$, and $K = (\Phi(X))^T\Phi(X)$ is the kernel matrix. The criterion given in (6) attains its maximum for the orthonormal vectors which satisfies

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T KWK\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T KK\boldsymbol{\alpha}} \tag{7}$$

There are many algorithms to find this optimal subspace and an orthonormal basis for it, for more information, one can refer to [4].

2.2. **Adaptive quasiconformal kernel.** In this part, we will present theoretically based AQKFD algorithm. As discussed in Subsection 2.1, kernel selection is a crucial problem for KFD, so we will use the so-called quasiconformal kernel to improve the performance of KFD. The quasiconformal kernel was first applied to improve SVM [8] and other kernel-based learning algorithms in the previous works [9,10]. We implement KFD in the high-dimensional quasiconformal kernel mapping space to develop AQKFD algorithm.

In AQKFD, Fisher criterion in the quasiconformal kernel mapping space is defined by

$$\tilde{J}_F^{\Phi}(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \tilde{K}W\tilde{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \tilde{K}\tilde{K}\boldsymbol{\alpha}} \tag{8}$$

where $\tilde{K}$ is the quasiconformal kernel matrix calculated by the quasiconformal kernel $\tilde{k}(x, y)$. The quasiconformal kernel is defined by

$$\tilde{k}(x, y) = q(x)q(y)k(x, y) \tag{9}$$

where $k(x, y)$ is a basic kernel such as a polynomial kernel or Gaussian kernel, $q(x)$ is a positive real valued function and different $q(x)$ make the quasiconformal kernel different properties, Pan and Li [10] expanded it by

$$q(x) = b_0 + \sum_{i=1}^{N_{XV}} b_i k_1(x, \tilde{x}_i) \tag{10}$$

where $k_1(x, \tilde{x}_i) = \exp(-\gamma_0||x - \tilde{x}_i||^2)$, $\tilde{x}_i$ $(i = 1, 2, \ldots, N_{XV})$ are "expansion vectors (XVs)" determined according to the distribution of the training data, N_{XV} is the number of XVs. b_i $(i = 1, 2, \ldots, N_{SV})$ are "expansion coefficients" associated with $\tilde{x}_i$, γ_0 is a free parameter.

It can be seen from Equations (9) and (10) that the quasiconformal kernel is determined by the "expansion coefficients" with the determinative "expansion vectors" and a free parameter. So the first thing is to choose all these procedural parameters. In this paper, XVs, which chosen to solve the expansion coefficients, are defined as follows [10]:

$$k_1(x, \tilde{x}_i) = k_1(x, \bar{x}_i) = \exp(-\gamma_0 ||x - \bar{x}_i||^2) \tag{11}$$

where $\bar{x}_i$ is the mean of each class.

Suppose given the free parameter γ_0 and the expansion vectors $\tilde{x}_i$ $(i = 1, 2, \ldots, N_{XV})$, we can create the matrix K_1 by

$$K_1 = \begin{bmatrix} 1 & k_1(x_1, \tilde{x}_1) & \cdots & k_1(x_1, \tilde{x}_{N_{SV}}) \\ 1 & k_1(x_2, \tilde{x}_1) & \cdots & k_1(x_2, \tilde{x}_{N_{SV}}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & k_1(x_N, \tilde{x}_1) & \cdots & k_1(x_N, \tilde{x}_{N_{SV}}) \end{bmatrix} \tag{12}$$

Let $\boldsymbol{\beta} = [b_0, b_1, \ldots, b_{N_{SV}}]^T$ and $Q = diag(q(x_1), q(x_2), \ldots, q(x_N))$, then we obtain

$$Q1_N = K_1\boldsymbol{\beta} \tag{13}$$

where 1_N is a N-dimensional vector whose entries are equal to 1.

Now, our purpose is to design an objective function which can find the adaptive expansion coefficients varied with the input data for the quasiconformal kernel. We firstly induce Xiong's method [9], explain its disadvantage and then propose ours. Xiong solves the expansion coefficient vector $\boldsymbol{\beta}$ through maximizing the measure of class separability of the data in the quasiconfromal kernel space via Fisher Criterion. The FC is defined as

$$J_F = \frac{tr(\tilde{S}_b^{\Phi})}{tr(\tilde{S}_t^{\Phi})} \tag{14}$$

where $\tilde{S}_b^{\Phi}$ and $\tilde{S}_t^{\Phi}$ denote the between-class scatter matrix and the total scatter matrix in the quasiconformal kernel mapping space which are similar to S_b^{Φ} and S_t^{Φ} defined in the feature space; tr and J_F denote the trace of a matrix and the Fisher scalar of measuring the class separabiltiy of the data.

In order to organize this paper conveniently, we lead to the following propositions:

Proposition 2.1. *Let K and $\tilde{K}$ denote the basic kernel matrix and quasiconformal kernel matrix respectively, then from Formula (9), we have $\tilde{K} = [q(x_i)q(x_j)k(x_i, x_j)]_{N\times N} = QKQ$.*

Proposition 2.2. *Assume $\tilde{K}_{ij}$ is a kernel matrix calculated with the ith and jth class mapped training samples and the kernel matrix, then $tr(\tilde{S}_b^{\Phi}) = 1_N^T\tilde{B}1_N$ and $tr(\tilde{S}_t^{\Phi}) = 1_N^T\tilde{T}1_N$, and $\tilde{B} = QBQ$, $\tilde{T} = QTQ$, where*

$$\tilde{B} = diag\left(\frac{1}{N_1}\tilde{K}_{11}, \ldots, \frac{1}{N_C}\tilde{K}_{CC}\right) - \frac{1}{N}\begin{bmatrix} \tilde{K}_{11} & \cdots & \tilde{K}_{1C} \\ \vdots & \ddots & \vdots \\ \tilde{K}_{C1} & \cdots & \tilde{K}_{CC} \end{bmatrix},$$

$$\tilde{T} = diag(\tilde{k}_{11}, \ldots, \tilde{k}_{NN}) - \frac{1}{N}\begin{bmatrix} \tilde{K}_{11} & \cdots & \tilde{K}_{1C} \\ \vdots & \ddots & \vdots \\ \tilde{K}_{C1} & \cdots & \tilde{K}_{CC} \end{bmatrix}$$

B and T are similar to $\tilde{B}$ and $\tilde{T}$ which are defined in the quasiconformal kernel mapping space.

Proof: Note the quasiconformal mapping Φ': X $\rightarrow$ Y, that is, $y_i = \Phi'(x_i)$. Let $Y = \{y_i^T | i = 1, 2, \ldots, N\}$, $Y_i = \{y_j^T | j = 1, 2, \ldots, N_i\}$, $i = 1, 2, \ldots, C$. Then, we have $\tilde{m}_i^{\Phi} = 1/N_i \sum_{j=1}^{N_i} y_j = 1/N_i \sum_{j=1}^{N_i} Y_i^T 1_{N_i}$ and $\tilde{m}_0^{\Phi} = 1/N \sum_{k=1}^{N} y_k = 1/N \sum_{k=1}^{N} Y^T 1_N$.

As the quasiconformal kernel space preserves the dot product, that is

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_C \end{bmatrix} \begin{bmatrix} Y_1^T & Y_2^T & \cdots & Y_C^T \end{bmatrix} = \begin{bmatrix} Y_1Y_1^T & Y_1Y_2^T & \cdots & Y_1Y_C^T \\ Y_2Y_1^T & Y_2Y_2^T & \cdots & Y_2Y_C^T \\ \vdots & \vdots & \ddots & \vdots \\ Y_CY_1^T & Y_CY_2^T & \cdots & Y_CY_C^T \end{bmatrix} = YY^T = \tilde{K}$$

$$= \begin{bmatrix} \tilde{K}_{11} & \tilde{K}_{12} & \ldots & \tilde{K}_{1C} \\ \tilde{K}_{21} & \tilde{K}_{22} & \ldots & \tilde{K}_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{K}_{C1} & \tilde{K}_{C2} & \ldots & \tilde{K}_{CC} \end{bmatrix}$$

Therefore,

$$\begin{aligned} tr(\tilde{S}_b^{\Phi}) &= \sum_{i=1}^{C} N_i (\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi})^T (\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi}) = \sum_{i=1}^{C} N_i (\tilde{m}_i^{\Phi})^T \tilde{m}_i^{\Phi} - (\tilde{m}_0^{\Phi})^T \tilde{m}_0^{\Phi} \\ &= \sum_{i=1}^{C} N_i \frac{1}{N_i} 1_{N_i}^T Y_i \frac{1}{N_i} Y_i^T 1_{N_i} - 1_N^T Y \frac{1}{N} Y^T 1_N \\ &= \sum_{i=1}^{C} \frac{1}{N_i} 1_{N_i}^T Y_i Y_i^T 1_{N_i} - \frac{1}{N} 1_N^T Y Y^T 1_N \\ &= \sum_{i=1}^{C} \frac{1}{N_i} 1_{N_i}^T \tilde{K}_{ii} 1_{N_i} - \frac{1}{N} 1_N^T \tilde{K} 1_N \\ &= 1_N^T \left\{ diag\left(\frac{1}{N_1} \tilde{K}_{11}, \ldots, \frac{1}{N_C} \tilde{K}_{CC} \right) - \frac{1}{N} \begin{bmatrix} \tilde{K}_{11} & \cdots & \tilde{K}_{1C} \\ \vdots & \ddots & \vdots \\ \tilde{K}_{C1} & \cdots & \tilde{K}_{CC} \end{bmatrix} \right\} 1_N \\ &= 1_N^T \tilde{B} 1_N \end{aligned}$$

$$\begin{aligned} tr(\tilde{S}_w^{\Phi}) &= \sum_{i=1}^{C} \sum_{j=1}^{N_i} (y_i^j - \tilde{m}_i^{\Phi})^T (y_i^j - \tilde{m}_i^{\Phi}) = \sum_{i=1}^{C} \left(\sum_{j=1}^{N_i} (y_i^j)^T y_i^j - N_i (\tilde{m}_i^{\Phi})^T \tilde{m}_i^{\Phi} \right) \\ &= \left(\sum_{k=1}^{N} y_k^T y_k - \sum_{i=1}^{C} N_i (\tilde{m}_i^{\Phi})^T \tilde{m}_i^{\Phi} \right) = \sum_{k=1}^{N} y_k^T y_k - \sum_{i=1}^{C} \frac{1}{N_i} 1_{N_i}^T \tilde{K}_{ii} 1_{N_i} \\ &= 1_N^T \left\{ diag(\tilde{k}_{11}, \ldots, \tilde{k}_{NN}) - diag\left(\frac{1}{N_1} \tilde{K}_{11}, \ldots, \frac{1}{N_C} \tilde{K}_{CC} \right) \right\} 1_N \end{aligned}$$

So,

$$\begin{aligned} tr(\tilde{S}_t^{\Phi}) &= tr(\tilde{S}_w^{\Phi}) + tr(\tilde{S}_b^{\Phi}) \\ &= 1_N^T \left\{ diag(\tilde{k}_{11}, \ldots, \tilde{k}_{NN}) - \frac{1}{N} \begin{bmatrix} \tilde{K}_{11} & \cdots & \tilde{K}_{1C} \\ \vdots & \ddots & \vdots \\ \tilde{K}_{C1} & \cdots & \tilde{K}_{CC} \end{bmatrix} \right\} 1_N \\ &= 1_N^T \tilde{T} 1_N. \end{aligned}$$

From Propositions 2.1 and 2.2, Fisher criterion in (14) is rewritten as

$$J_F = \frac{tr(\tilde{S}_b^{\Phi})}{tr(\tilde{S}_t^{\Phi})} = \frac{1_N^T \tilde{B} 1_N}{1_N^T \tilde{T} 1_N} = \frac{1_N^T Q B Q 1_N}{1_N^T Q T Q 1_N} = \frac{\boldsymbol{\beta}^T K_1^T B K_1 \boldsymbol{\beta}}{\boldsymbol{\beta}^T K_1^T T K_1 \boldsymbol{\beta}} \tag{15}$$

Denote $\widehat{B} = K_1^T B K_1$ and $\widehat{T} = K_1^T T K_1$, then (15) is given by

$$\tilde{J}_F(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^T \widehat{B} \boldsymbol{\beta}}{\boldsymbol{\beta}^T \widehat{T} \boldsymbol{\beta}} \tag{16}$$

As $\widehat{B}$ and $\widehat{T}$ are the constant matrices, meanwhile, XVs and the free parameter are determined in Subsection 2.1, then $\tilde{J}_F(\boldsymbol{\beta})$ in (16) is a function with its variable $\boldsymbol{\beta}$. Thus, it can be transformed into an objective function constrained by the unit vector $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$, to maximize $\tilde{J}_F(\boldsymbol{\beta})$, which can be described as

$$\max \quad \frac{\boldsymbol{\beta}^T \widehat{B} \boldsymbol{\beta}}{\boldsymbol{\beta}^T \widehat{T} \boldsymbol{\beta}} \quad \text{s.t.} \quad \boldsymbol{\beta}^T\boldsymbol{\beta} - 1 = 0 \tag{17}$$

Considering that $\widehat{T}$ may be a singular matrix, Xiong obtained the optimal discriminant vector $\boldsymbol{\beta}^*$ by iteratively updating the algorithm. The limitation of this method lies in the difficulty in finding the unique optimal solution and the high time-consuming. Therefore, in order to reduce computation time and solve the singular problem, we propose another criterion called weighted maximum margin to seek optimal solution $\boldsymbol{\beta}^*$. Detailed information is discussed in Section 3.

3. Weighted Maximum Margin Criterion (WMMC).

In this section, we introduce the weighted maximum margin criterion and solve the parameters of adaptive quasiconformal kernel through it. Simultaneously, we will illustrate that the best weight parameter of WMMC is just the optimal solution of Equation (16) which is discussed in Subsection 2.2.

3.1. WMMC.

Maximum margin criterion is first proposed by Li [11] in order to extract the feature by maximizing the average margin between different classes of data in the feature space. Pan and Li [11] use it to measure the average margin between different classes of data in the quasiconformal kernel mapping space, the average margin between two classes C_i and C_j in the quasiconformal kernel mapping space is defined as

$$\tilde{J}_W = \frac{1}{2N} \sum_{i=1}^{C} \sum_{j=1}^{C} N_i N_j d(C_i, C_j) \tag{18}$$

Li and Pan define the margin as

$$d(C_i, C_j) = d(\tilde{m}_i^{\Phi}, \tilde{m}_j^{\Phi}) - (s(C_i) + s(C_j)) \tag{19}$$

where $d(m_i, m_j)$ means the distance between mean vectors of the class C_i and C_j; $s(C_i)$ and $s(C_j)$ denote the measure of the scatter of the class C_i and C_j respectively. The authors use the within-class scatter matrix of class i and j to represent $s(C_i)$ and $s(C_j)$, that is, $s(C_i) = tr(\tilde{S}_i^{\Phi})$ and $s(C_j) = tr(\tilde{S}_j^{\Phi})$.

However, the weakness of the above representation is that it does not take the differences of within-class average distance and between-class average distances into account. It means that we should set parameters to weigh the two different kinds of distances. In this paper, we define it as follows:

$$d(C_i, C_j) = (1-t) d(\tilde{m}_i^{\Phi}, \tilde{m}_j^{\Phi}) - t(tr(\tilde{S}_i^{\Phi}) + tr(\tilde{S}_j^{\Phi})) \tag{20}$$

where t is a variable and $0 < t < 1$.

Proposition 3.1. *Let $tr(\tilde{S}_b^{\Phi})$ and $tr(\tilde{S}_t^{\Phi})$ denote the trace of between-class scatter matrix and the total scatter matrix which are defined in Subsection 2.2, then $\tilde{J}_W = tr(\tilde{S}_b^{\Phi}) - t \cdot tr(\tilde{S}_t^{\Phi})$.*

Proof:

$$\tilde{J}_W = \frac{1}{2N}\sum_{i=1}^{C}\sum_{j=1}^{C} N_i N_j d(C_i, C_j) = \frac{1}{2N}\sum_{i=1}^{C}\sum_{j=1}^{C} N_i N_j((1-t)\cdot d(m_i, m_j) - t(tr(\tilde{S}_i^{\Phi}) + tr(\tilde{S}_j^{\Phi})))$$

Denote $p_i = N_i/N$, then $tr(\tilde{S}_b^{\Phi}) = \sum_{i=1}^{C} N_i(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi})^T(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi}) = N\sum_{i=1}^{C} p_i(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi})^T(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi}) tr(\tilde{S}_w^{\Phi}) = \sum_{i=1}^{C}\sum_{j=1}^{N_i}(y_i^j - \tilde{m}_i^{\Phi})^T(y_i^j - \tilde{m}_i^{\Phi}) = N\sum_{i=1}^{C} p_i tr(\tilde{S}_i^{\Phi})$.

In order to calculate conveniently, decompose $\tilde{J}_W$ into two parts:

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^{C}\sum_{j=1}^{C} N_i N_j (1-t) d(\tilde{m}_i^{\Phi}, \tilde{m}_j^{\Phi}) \\
&= \frac{(1-t)N}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j d(\tilde{m}_i^{\Phi}, \tilde{m}_j^{\Phi}) \\
&= \frac{(1-t)N}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\tilde{m}_i^{\Phi} - \tilde{m}_j^{\Phi})^T(\tilde{m}_i^{\Phi} - \tilde{m}_j^{\Phi}) \\
&= \frac{(1-t)N}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi} + \tilde{m}_0^{\Phi} - \tilde{m}_j^{\Phi})^T(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi} + \tilde{m}_0^{\Phi} - \tilde{m}_j^{\Phi})
\end{aligned}$$

We can simplify the above equation to $\sum_{i=1}^{C} p_i(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi})^T(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi})$ by using the fact that $\sum_{i=1}^{C} p_i(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi}) = 0$. So,

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^{C}\sum_{j=1}^{C} N_i N_j (1-t) d(\tilde{m}_i^{\Phi}, \tilde{m}_j^{\Phi}) \\
&= \frac{(1-t)}{2} N \sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j d(\tilde{m}_i^{\Phi}, \tilde{m}_j^{\Phi}) \\
&= (1-t)N\sum_{i=1}^{C} p_i(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi})^T(\tilde{m}_i^{\Phi} - \tilde{m}_0^{\Phi}) \\
&= (1-t)\cdot tr(\tilde{S}_b^{\Phi}) \\
&\frac{1}{2N}\sum_{i=1}^{C}\sum_{j=1}^{C} N_i N_j t(tr(\tilde{S}_i^{\Phi}) + tr(\tilde{S}_j^{\Phi})) \\
&= \frac{tN}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (tr(\tilde{S}_i^{\Phi}) + tr(\tilde{S}_j^{\Phi})) \\
&= tN\sum_{i=1}^{C} p_i tr(\tilde{S}_i^{\Phi}) = t\cdot tr(\tilde{S}_w^{\Phi})
\end{aligned}$$

Therefore, $\tilde{J}_W = (1-t)\cdot tr(\tilde{S}_b^{\Phi}) - t\cdot tr(\tilde{S}_w^{\Phi}) = tr(\tilde{S}_b^{\Phi}) - t\cdot tr(\tilde{S}_t^{\Phi})$.

From Propositions 2.2 and 3.1, we obtain

$$\tilde{J}_W = 1_N^T(\tilde{B} - t\cdot\tilde{T})1_N = \boldsymbol{\beta}^T K_1^T(B - t\cdot T)K_1\boldsymbol{\beta} = \boldsymbol{\beta}^T(\widehat{B} - t\cdot\widehat{T})\boldsymbol{\beta} = \tilde{J}_W(\boldsymbol{\beta}) \tag{21}$$

Equation (21) is called weighted maximum margin criterion (WMMC).

Thus, an objective function constrained by the unit vector $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$, is defined as

$$\max \quad \boldsymbol{\beta}^T(\widehat{B} - t\cdot\widehat{T})\boldsymbol{\beta} \quad \text{s.t.} \quad \boldsymbol{\beta}^T\boldsymbol{\beta} - 1 = 0 \tag{22}$$

The solution of the above constrained optimization problem is the so-called Lagrangian method, and it can be transformed into the following eigenvalue equation with parameter λ.

$$(\widehat{B} - t\cdot\widehat{T})\boldsymbol{\beta} = \lambda\boldsymbol{\beta} \tag{23}$$

Multiply $\boldsymbol{\beta}^T$ at left side of Equation (23), then

$$\boldsymbol{\beta}^T(\widehat{B} - t\cdot\widehat{T})\boldsymbol{\beta} = \boldsymbol{\beta}^T\lambda\boldsymbol{\beta} = \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} = \lambda \tag{24}$$

Finding the optimal expansion coefficient vector $\boldsymbol{\beta}^*$ is equal to solve the eigenvector of $(\widehat{B} - t\cdot\widehat{T})$ corresponding to its largest eigenvalue λ. From (21) and (24), we obtain

$$\tilde{J}_W = \lambda \tag{25}$$

Therefore, it means that to maximize the average margin $\tilde{J}_W$ is equivalent to work out λ. As to how to set the weight parameter t, we will discuss it in the following section.

3.2. **The selection of the weight parameter.** It is known that in order to improve the discriminant ability of AQKFD, we expect the optimal vector $\boldsymbol{\beta}^*$ achieves larger values of Fisher discriminant criterion. The key of doing this is the selection of the weight parameter. In this section, we will discuss the problem on how to choose the weight parameter t.

Let ω^*, φ^* be the first discriminant vectors of (16) and (21) respectively. Then we have

$$\tilde{J}_F(\omega^*) = \frac{(\omega^*)^T\widehat{B}\omega^*}{(\omega^*)^T\widehat{T}\omega^*} = \max_{\boldsymbol{\beta}} \tilde{J}_F(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^T\widehat{B}\boldsymbol{\beta}}{\boldsymbol{\beta}^T\widehat{T}\boldsymbol{\beta}} \tag{26}$$

$$\tilde{J}_W(\varphi^*) = (\varphi^*)^T(\widehat{B} - t\cdot\widehat{T})\varphi^* = \max_{\boldsymbol{\beta}} \tilde{J}_W(\boldsymbol{\beta}) = \boldsymbol{\beta}^T(\widehat{B} - t\cdot\widehat{T})\boldsymbol{\beta} \tag{27}$$

Equation (26) can be written in the following form:

$$(\omega^*)^T(\widehat{B} - \tilde{J}_F(\omega^*)\cdot\widehat{T})\omega^* = 0 \tag{28}$$

Then, we have

$$\begin{aligned}(\omega^*)^T(\widehat{B} - t\cdot\widehat{T})\omega^* &= (\omega^*)^T(\widehat{B} - \tilde{J}_F(\omega^*)\cdot\widehat{T} + \tilde{J}_F(\omega^*)\cdot\widehat{T} - t\widehat{T})\omega^* \\ &= (\omega^*)^T(\widehat{B} - \tilde{J}_F(\omega^*)\cdot\widehat{T})\omega^* + (\omega^*)^T(\tilde{J}_F(\omega^*) - t)\widehat{T}\omega^* = (\tilde{J}_F(\omega^*) - t)(\omega^*)^T\widehat{T}\omega^*\end{aligned} \tag{29}$$

In addition, from Equation (27)

$$\tilde{J}_W(\varphi^*) = \max_{\boldsymbol{\beta}} \tilde{J}_W(\boldsymbol{\beta}) \geq \tilde{J}_W(\omega^*) = (\omega^*)^T(\widehat{B} - t\cdot\widehat{T})\omega^* > 0 \tag{30}$$

Thus, from Equations (26) and (30), we have

$$\frac{(\varphi^*)^T\widehat{B}\varphi^*}{(\varphi^*)^T\widehat{T}\varphi^*} = \frac{\tilde{J}_W(\varphi^*)}{(\varphi^*)^T\widehat{T}\varphi^*} + t > t \tag{31}$$

Moreover,

$$\tilde{J}_F(\omega^*) = \max_{\boldsymbol{\beta}} \tilde{J}_F(\boldsymbol{\beta}) \geq \tilde{J}_F(\varphi^*) = \frac{(\varphi^*)^T \widehat{B} \varphi^*}{(\varphi^*)^T \widehat{T} \varphi^*} \tag{32}$$

From Equations (31) and (32), we obtain that

$$t < \frac{(\varphi^*)^T \widehat{B} \varphi^*}{(\varphi^*)^T \widehat{T} \varphi^*} < \tilde{J}_F(\omega^*) \tag{33}$$

Therefore, it means that when t is approximate to $\tilde{J}_F(\omega^*)$,

$$\tilde{J}_F(\varphi^*) = \frac{(\varphi^*)^T \widehat{B} \varphi^*}{(\varphi^*)^T \widehat{T} \varphi^*} \to \tilde{J}_F(\omega^*) = \max_{\boldsymbol{\beta}} \tilde{J}_F(\boldsymbol{\beta}) \tag{34}$$

Thus far, we have settled how to determine the weight parameter t. In other words, if we expect that the optimal discriminant vectors of WMMC can extract the feature by maximizing the average margin between different classes in terms of the Fisher's discriminant criterion, we might choose a larger weight parameter. In this paper, we empirically set the weight parameter t to 0.7.

3.3. **AQKFD-WMMC algorithm.** The objective function designed based on FC of (8) and WMMC of (21) is to maximize the class separability of the data in the optimal quasiconformal kernel mapping space. After solving the adaptive parameters for the quasiconformal kernel via WMMC, we implement KFD algorithm in the optimal quasiconformal kernel mapping space with feature extraction shown in Equation (8). The optimal projection α^* from the feature space to the projection subspace is calculated by the same way as KFD. The complete AQKDA-WMMC algorithm is given below:

Input: A set of n-dimensional training samples $\{x_1, x_2, \ldots, x_N\}$.

Output: A lower m-dimensional feature representation y of x.

Step 1. Set basic kernel K, and calculate matrix K_1, B, T, $\widehat{B}$, $\widehat{T}$.
Step 2. Set weight parameter t, and work out $\boldsymbol{\beta}^*$ via WMMC in Equation (21).
Step 3. Calculate the quasiconformal kernel $\tilde{K}$ with $\boldsymbol{\beta}^*$.
Step 4. Obtain the discriminant projection matrix $A = [\alpha_1, \alpha_2, \ldots, \alpha_d]$ by solving Equation (8).
Step 5. Extract feature vector of sample x with $y = A^T[\tilde{k}(x_1, x), \tilde{k}(x_2, x), \ldots, \tilde{k}(x_N, x)]$.

4. **Experiments Results.** In this section, the performance of AQKFD-WMMC is evaluated on three face databases, i.e., ORL face database, Yale face database and FERET face database. Firstly, we evaluated the proposed method compared with AQKFD-FC [9] on time-consuming and recognition rate. Secondly, we test the superiority of the proposed algorithm compared with KFD, KPCA, AQKFD-MMC [10] and KWMMDA [12].

4.1. **Face dataset and experiment description.** The ORL face database, developed at the Olivetti Research Laboratory, Cambridge, U.K., is composed of 400 gray-scale images with 10 images for each of 40 individuals. The variations of the images are across pose, time and facial expression. The Yale face database was constructed at the Yale Center for Computational Vision and Control. It contains 165 gray-scale images of 15 individuals. These images are taken under different lighting condition (left-light, center-light and right-light), and different facial expression (normal, happy, sad, sleepy, surprised and wink), and with/without glasses. The FERET face database has become a standard database for testing and evaluating state-of-the-art face recognition algorithms. It is

composed of 1, 400 images of 200 individuals (each one has seven images) and it involves variations in facial expression, illumination, and pose.

In our experiments, to reduce computation complexity, we resize the original ORL face images sized 112×92 pixels with a 256 gray scale to 48×48 pixels, and examples are shown in Figure 1(a). Similarly, the images from Yale databases are cropped to the size of 100×100 pixels, and some examples are shown in Figure 1(b) and the images from FERET database were resized to 40×40 pixels and further preprocessed by histogram equalization and some images of one person are shown in Figure 1(c).

(a) Example cropped face images from the ORL face database (cropped to the size of 48×48 to extract the facial region)

(b) Example cropped face images from the Yale face database (cropped to the size of 100×100 to extract the facial region)

(c) Example cropped face images from the FERET database (cropped to the size of 80×80 to extract the facial region)

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

After describing the databases used in our experiments, it is worthwhile to make some remarks on the experiment setting as follows. (1) We randomly select 5 images from each subject from ORL face database for training, and the rest images are used to test the performance. Similarly, 6 (and 4) images of each person randomly selected from YALE (and FERET) database are used to construct the training set, and the rest images of each person are used to test the performance of the algorithms. (2) We run each set of experiments for 10 times and the averaged results are used to evaluate the performance of the proposed algorithm. Besides, the procedural parameters are chosen with cross-validation method in the second experiment allowing for the kernel parameters. (3) The experiments are implemented on a AMD 2.40GHz computer with 4.00GB RAM and programmed in a MATLAB 2010a platform. (4) The free parameter of quasiconformal kernel is and the dimension of the quasiconformal kernel space is 40.

4.2. **Recognition results on ORL, YALE and FERET datasets.** In this section, the experiment is divided into two parts. Here, notice that, for the polynomial kernel $k(x, y) = (x \cdot y)^d$ $(d \in N)$, the candidate order interval is from 1 to 6, and the parameter of Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/\sigma)$ is from 1e5 to 1e10. Moreover, a nearest neighbor classifier (NN) is employed in the whole experiment.

Firstly, AQKFD-FC and the proposed AQKFD-WMMC are tested and compared. In order to gain more insights, time-consuming and recognition rate are both considered as the factors to evaluate the performance. The results are presented from Table 1 to Table 6, as it shows, AQKFD-WMMC can give higher recognition accuracy and lower time-consuming compared with AQKFD-FC. The result can be explained as follows.

Table 1. AQKFD-FC vs. AQKFD-WMMC on ORL using Polynomial kernel

	Polynomial(d)	1	2	3	4	5	6
Recog rate (%)	AQKFD-FC	93.00	91.50	90.50	88.00	82.50	77.50
	AQKFD-WMMC	96.00	94.50	93.00	90.50	88.00	85.50
Time-consuming (s)	AQKFD-FC	0.55	0.19	0.33	0.51	0.51	0.42
	AQKFD-WMMC	0.01	0.01	0.01	0.01	0.01	0.01

Table 2. AQKFD-FC vs. AQKFD-WMMC on ORL using Gaussian kernel

	Gaussian(σ)	1e5	1e6	1e7	1e8	1e9	1e10
Recog rate (%)	AQKFD-FC	82.00	84.00	94.00	93.00	95.00	91.00
	AQKFD-WMMC	85.00	88.00	94.00	96.00	96.50	94.50
Time-consuming (s)	AQKFD-FC	0.35	0.37	0.23	0.65	0.42	0.37
	AQKFD-WMMC	0.03	0.03	0.02	0.04	0.03	0.03

Table 3. AQKFD-FC vs. AQKFD-WMMC on YALE using Polynomial kernel

	Polynomial(d)	1	2	3	4	5	6
Recog rate (%)	AQKFD-FC	89.33	85.33	84.00	81.33	76.00	73.33
	AQKFD-WMMC	93.33	90.67	88.00	85.33	82.67	81.33
Time-consuming (s)	AQKFD-FC	0.19	0.16	0.25	0.25	0.12	0.18
	AQKFD-WMMC	0.01	0.01	0.01	0.01	0.01	0.01

Table 4. AQKFD-FC vs. AQKFD-WMMC on YALE using Gaussian kernel

	Gaussian(σ)	1e5	1e6	1e7	1e8	1e9	1e10
Recog rate (%)	AQKFD-FC	73.33	76.00	77.33	80.00	85.33	81.33
	AQKFD-WMMC	74.67	77.33	84.00	89.33	92.00	85.33
Time-consuming (s)	AQKFD-FC	0.25	0.11	0.25	0.19	0.25	0.16
	AQKFD-WMMC	0.06	0.02	0.06	0.06	0.03	0.06

Since AQKFD-WMMC leads to the largest class discrimination in quasiconformal kernel feature space via weighted maximum margin criterion of the original input data, it is obviously lower time-consuming than that of AQKFD-FC as it takes the Fisher criterion into consideration. Moreover, our proposed AQKFD-WMMC is a self-adaption method. It means no matter what the original input data is, we can always seek the optimal quasiconformal kernel feature space using AQKFD and the optimal selection of the weight parameter using WMMC in the theoretical-based way. However, AQKFD-FC is a suboptimal way, because Fisher criterion incurs the matrix singular problem when solving the eigenvalue problem. So, it adopts an updating algorithm, which can just get the suboptimal result. Therefore, AQKFD-WMMC can gain higher recognition rate than AQKFD-FC as well as lower time-consuming.

Secondly, we demonstrate the superiority of AQKFD-WMMC compared with KFD, KPCA, AQKFD-MMC and KWMMDA. All the procedural parameters of these algorithms are chosen with cross-validation methods. As results shown in Table 7, AQKFD-WMMC achieves the highest recognition rate compared with other algorithms.

The reasonable explanations are given as follows:

Firstly, as the conventional KFD and KPCA methods do not optimize the kernel in the feature space, so these two gain the worst recognition accuracy when compared with

TABLE 5. AQKFD-FC vs. AQKFD-WMMC on FERET using Polynomial kernel

	Polynomial(d)	1	2	3	4	5	6
Recog rate (%)	AQKFD-FC	86.83	85.67	85.50	82.00	80.50	76.50
	AQKFD-WMMC	88.33	88.33	88.17	86.67	83.67	80.33
Time-consuming (s)	AQKFD-FC	1.50	0.89	0.93	1.51	1.60	1.22
	AQKFD-WMMC	0.03	0.04	0.03	0.04	0.06	0.04

TABLE 6. AQKFD-FC vs. AQKFD-WMMC on FERET using Gaussian kernel

	Gaussian(σ)	1e5	1e6	1e7	1e8	1e9	1e10
Recog rate (%)	AQKFD-FC	76.00	80.33	83.33	84.33	86.00	83.00
	AQKFD-WMMC	80.33	82.00	84.67	87.33	88.00	85.50
Time-consuming (s)	AQKFD-FC	1.15	1.27	0.73	1.55	1.42	1.37
	AQKFD-WMMC	0.10	0.08	0.07	0.11	0.09	0.09

TABLE 7. Average recognition rates (%) of AQKFD-WMMC, KFD, KPCA, AQKFD-MMC, KWMMDA across 10 tests on ORL, YALE and FERET datasets

	ORL	YALE	FERET
AQKFD-WMMC	93.50	85.10	84.67
KFD	90.00	74.67	75.33
KPCA	84.80	77.89	76.00
AQKFD-MMC	91.38	83.00	81.67
KWMMDA	87.13	75.78	72.50

other three methods. Secondly, since the optimal discriminant vectors of AQKFD-WMMC contain better discriminant information than AQKFD-MMC in terms of the Fisher discriminant criterion, AQKFD-WMMC could obtain better results than AQKFD-MMC. In fact, AQKFD-MMC is the special case of AQKFD-WMMC when the weigh parameter is 0.5. Thirdly, as mentioned above, AQKFD-WMMC is a self-adaption method. By using AQKFD, it can seek the optimal quasiconformal kernel, so it is more adaptive to the original input data for classification than that of KWMMDA, and thus it has higher recognition rate than KWMMDA.

From the above experiments and explanations, we can come to the following conclusions: (1) Experiment results indicate that AQKFD-WMMC consistently performs the best among all these algorithms. (2) By comparing AQKFD-WMMC and AQKFD-MMC, experimental result shows that AQKFD-WMMC could obtain better results than AQKFD-MMC, because the optimal discriminant vectors of AQKFD-WMMC contain better discriminant information than AQKFD-MMC in terms of the Fisher discriminant criterion. (3) By comparing AQKFD-WMMC and KWMMDA, the result demonstrates the fact that the quasiconformal kernel of AQKFD-WMMC is indeed important as it is adaptive to the input data for classification. (4) Taking two criterions of solving the expansion coefficients, i.e., Fisher criterion and WMMC into account, WMMC can solve the singular problem which is endured by the FC. Besides, it is lower time-consuming when compared with FC.

5. **Conclusion.** In this paper, we apply a quasiconformal kernel and propose a weighted maximum margin criterion to develop a kernel-based learning method called AQKFD-WMMC for extracting nonlinear features. AQKFD-WMMC is more adaptive to the input data for classification because it changes the kernel structure automatically with the adaptive parameters which are computed through maximizing the measure of class separability of the data in the quasiconformal kernel mapping space. Moreover, AQKFD-WMMC can obtain better result since the optimal discriminant vectors of AQKFD-WMMC contain better discriminant information in terms of Fisher discriminant criterion. Besides, AQKFD-WMMC avoid the matrix singular problem and it is low time-consuming which is endured by KFD.

Acknowledgment. This work is partially supported by Major Program of Natural Science Foundation of China, No. 61033012, Natural Science Foundation of China, No. 6110 03177, and Fundamental Research Funds for the Central Universities, No. 1600-852016 and No. DUT12JR07. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
[2] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda and B. Scholkopf, An introduction to kernel based learning algorithms, *IEEE Trans. Neural Networks*, vol.12, no.2, pp.181-201, 2001.
[3] B. Scholkopf, A. Smola and K. R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, vol.10, no.5, pp.1299-1319, 1998.
[4] S. Mika, G. Rätsch, J. Weston, B. Schölkopf and K.-R. Müller, Fisher discriminant analysis with kernels, *Proc. of IEEE the 9th Int'l Workshop Neural Networks for Signal Processing*, pp.41-48, 1999.
[5] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola and K.-R. Müller, Invariant feature extraction and classification in kernel spaces, *Proc. of the 12th Conference on Advances in Neural Information Processing Systems*, 1999.
[6] J. Lu, K. N. Plataniotis and A. N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, *IEEE Trans. Neural Networks*, vol.14, no.1, pp.117-226, 2003.
[7] G. Baudat and F. Anouar, Generalized discriminant analysis using a kernel approach, *Neural Computation*, vol.12, no.10, pp.2385-2404, 2000.
[8] Z. Liang and P. Shi, Efficient algorithm for kernel discriminant anlaysis, *Pattern Recognition*, vol.37, no.2, pp.381-384, 2000.
[9] Z. Liang and P. Shi, An efficient and effective method to solve kernel Fisher discriminant analysis, *Neurocomputing*, vol.61, pp.485-493, 2004.
[10] Z. Liang and P. Shi, Uncorrelated discriminant vectors using a kernel method, *Pattern Recognition*, vol.38, pp.307-310, 2005.
[11] J. Yang, A. F. Frangi, J.-Y. Yang, D. Zhang and Z. Jin, KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.27, no.2, pp.230-244, 2005.
[12] M. H. Yang, Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods, *Proc. of the 5th IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pp.215-220, 2002.
[13] Y. I. Zheng, J. Yang, J.-Y. Yang and X.-J. Wu, A reformative kernel fisher discriminant algorithm and its application to face recognition, *Neurocomputing*, vol.69, no.13, pp.1806-1810, 2006.
[14] J. Huang, P. C. Yuen, W.-S. Chen and J. H. Lai, Kernel subspace LDA with optimized kernel parameters on face recognition, *Proc. of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.
[15] L. Wang, K. L. Chan and P. Xue, A criterion for optimizing kernel parameters in KBDA for image retrieval, *IEEE Trans. Systems, Man Cybernet B: Cybernet.*, vol.35, no.2, pp.556-562, 2005.
[16] B. Ma, H.-Y. Qu and H.-S. Wong, Kernel clustering-based discriminant analysis, *Pattern Recognition*, vol.40, no.1, pp.324-327, 2007.
[17] W.-S. Chen, P. C. Yuen and J. Huang, Kernel machine-based one-parameter regularized Fisher discriminant method for face recognition, *IEEE Trans. Systems, Man Cybernet B: Cybernet.*, vol.35, no.4, pp.658-669, 2005.

[18] Y. Liang, C. Li, W. Gong and Y. Pan, Uncorrelated linear discriminant analysis based on weighted pairwise Fisher criterion, *Pattern Recognition*, vol.40, pp.3606-3615, 2007.
[19] D. Tao, X. Tang, X. Li and Y. Rui, Direct kernel biased discriminant analysis: A new content based image retrieval relevance feedback algorithm, *IEEE Trans. Multimedia*, vol.8, no.4, pp.716-727, 2006.
[20] Y. Xu, D. Zhang, Z. Jin, M. Li and J.-Y. Yang, A fast kernel-based nonlinear discriminant analysis for multi-class problems, *Pattern Recognition*, vol.39, no.6, pp.1026-1033, 2006.
[21] D.-Y. Yeung, H. Chang and G. Dai, Learning the kernel matrix by maximizing a KFD-based class separability criterion, *Pattern Recognition*, vol.40, no.7, pp.2021-2028, 2007.
[22] K. Saadi, N. L. C. Talbot and G. C. Cawley, Optimally regularised kernel Fisher discriminant classification, *Neural Networks*, vol.20, no.7, pp.832-841, 2007.
[23] Q. Liu, H. Lu and S. Ma, Improving kernel Fisher discriminant analysis for face recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.14, no.1, pp.42-49, 2004.
[24] L. Shen, L. Bai and M. Fairhurst, Gabor wavelets and general discriminant analysis for face identification and verification, *Image Vis. Comput.*, vol.25, no.5, pp.553-563, 2007.
[25] X.-H. Wu and J.-J. Zhou, Fuzzy discriminant analysis with kernel methods, *Pattern Recognition*, vol.39, no.11, pp.2236-2239, 2006.
[26] N. Cristianini, J. Kandola, A. Elisseeff and J. Shawe-Taylor, On kernel target alignment, *Proc. of Neural Information Processing Systems*, pp.367-373, 2001.
[27] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui and M. I. Jordan, Learning the kernel matrix with semidefinte programming, *J. Mach. Learn. Res.*, vol.5, 2004.
[28] S. Amari and S. Wu, Improving support vector machine classifiers by modifying kernel functions, *Neural Network*, vol.12, no.6, pp.783-789, 1999.
[29] H. Xiong, M. N. S. Swamy and M. O. Ahmad, Optimizing the kernel in the empirical feature space, *IEEE Trans. Neural Networks*, vol.16, no.2, pp.460-474, 2005.
[30] J.-S. Pan, J.-B. Li and Z.-M. Lu, Adaptive quasiconformal kernel discriminant analysis, *Neurocomputing*, vol.71, pp.2754-2760, 2008.
[31] H. Li, T. Jiang and K. Zhang, Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Networks*, vol.17, no.1, pp.157-165, 2006.
[32] W. Zheng, C. Zou and L. Zhao, Weighted maximum margin discriminant analysis with kernels, *Pattern Recognition*, vol.67, pp.357-362, 2005.
[33] S.-J. Kim, A. Magnani and S. Boyd, Optimal kernel selection in kernel Fisher discriminant analysis, *Proc. of the 23rd International Conference on Machine Learning*, 2006.
[34] O. C. Hamsici and A. M. Martinez, Sparse kernels for bayes optimal discriminant analysis, *Proc. of CVPR*, 2009.
[35] D. You, O. C. Hamsici and A. M. Martinez, Kernel optimization in discriminant analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.33, no.3, pp.631-639, 2011.