Foley–Sammon Optimal Discriminant Vectors Using Kernel Approach

Wenming Zheng, Li Zhao, and Cairong Zou, *Member, IEEE*

***Abstract*—A new nonlinear feature extraction method called kernel Foley–Sammon optimal discriminant vectors (KFSODVs) is presented in this paper. This new method extends the well-known Foley–Sammon optimal discriminant vectors (FSODVs) from linear domain to a nonlinear domain via the kernel trick that has been used in support vector machine (SVM) and other commonly used kernel-based learning algorithms. The proposed method also provides an effective technique to solve the so-called small sample size (SSS) problem which exists in many classification problems such as face recognition. We give the derivation of KFSODV and conduct experiments on both simulated and real data sets to confirm that the KFSODV method is superior to the previous commonly used kernel-based learning algorithms in terms of the performance of discrimination.**

***Index Terms*—Face recognition, Foley–Sammon optimal discriminant vectors (FSODVs), kernel methods, kernel principal component analysis (PCA), null space.**

I. Introduction

FEATURE EXTRACTION plays a very important role in pattern recognition. Up to now, a lot of feature extraction approaches have been proposed to this goal. Among these, the principal component analysis (PCA) [7] and the Fisher linear discriminant analysis (LDA or FLD) [4], [14] are the two most successful ones. PCA is an orthogonal basis transformation for extracting the structure from possibly high-dimensional data sets [10], whereas LDA is a powerful technique used for extracting the discriminant features from the data sets. As for pattern classification, however, it was shown that the LDA-based approach outperforms the PCA-based approach [19]. The former realizes an optimal feature space on the Fisher's criterion function, subject to the uncorrelated constraints of the discriminant vectors. Nevertheless, the classification performance of LDA is still limited because the number of discriminant vectors computed by LDA generally depends on the number of the pattern classes [3]. To overcome this drawback, Sammon [1] proposed an optimal discriminant plane (ODP) under the orthogonal constraints of the discriminant vectors in 1970. In 1975, Foley and Sammon [2] extended the ODP method and presented the well-known Foley–Sammon optimal discriminant vectors (FSODVs), which can obtain much more discriminant vectors than LDA in most cases. Although successful to many linear patterns, FSODV fails to deliver good performance when it comes to the nonlinear patterns such as face patterns, which subject to large variations in viewpoints, resulting in a highly nonconvex and complex distribution [18]. Thus, it is necessary to extend the FSODV method to be suitable for nonlinear patterns.

Motivated by support vector machine (SVM) [11], kernel principal component analysis (KPCA) [10], kernel fisher discriminant analysis (KFD) [16], and generalized discriminant analysis (GDA) [9], we developed a nonlinear FSODV approach via the kernel trick in this paper, kernel Foley–Sammon optimal discriminant vectors (KFSODVs). The KFSODV method combines the strengths of both FSODV and kernel-based learning techniques [12], [13] to improve the performance of FSODV. Besides, the proposed algorithm also provides an effective technique to solve the so-called small sample size (SSS) problem [8], where the dimensionality of the feature space is always larger than the number of the training samples, such that the most discriminant solutions of KFSODV lie in the null space of the within-class matrix [21], [22]. We will give detailed derivation of the formulations of KFSODV and also make comparison to other commonly used kernel-based learning algorithms on both theoretical and experimental analysis in this paper.

The remainder of this paper is organized as follows: In Section II, we introduce the KFSODV method and some related theorems. In Section III, we derive the formulations of KFSODV. Experiments are conducted on both simulated and real data sets in Section IV. Discussion and conclusion are given in Section V.

II. Related Work

A. FSODVs in Feature Space

Assume that $\mathbf{X}$ is a n-dimensional sample set with N elements belonging to c classes. Let $\mathbf{X}_l$ denote the lth class sample set, and N_l the cardinality of $\mathbf{X}_l$. Thus, we have

$$\mathbf{X} = \bigcup_{l=1}^{c} \mathbf{X}_l, \quad N = \sum_{l=1}^{c} N_l.$$

The between-class scatter matrix $\mathbf{S_B}$, the within-class scatter matrix $\mathbf{S_W}$ and the total-class scatter matrix $\mathbf{S_T}$ are defined as follows:

$$\begin{aligned}
\mathbf{S_B} &= \sum_{i=1}^{c} N_i(\mathbf{u}_i - \mathbf{u})(\mathbf{u}_i - \mathbf{u})^T \\
\mathbf{S_W} &= \sum_{i=1}^{c}\sum_{j=1}^{N_i} \left(\mathbf{x}_i^j - \mathbf{u}_i\right)\left(\mathbf{x}_i^j - \mathbf{u}_i\right)^T \\
\mathbf{S_T} &= \sum_{i=1}^{c}\sum_{j=1}^{N_i} \left(\mathbf{x}_i^j - \mathbf{u}\right)\left(\mathbf{x}_i^j - \mathbf{u}\right)^T
\end{aligned} \tag{1}$$

Manuscript received February 9, 2003; revised September 16, 2003.

The authors are with the Engineering Research Center of Information Processing and Application, Southeast University, Nanjing, Jiangsu 210096, China (e-mail: wenming_zheng@seu.edu.cn).

Digital Object Identifier 10.1109/TNN.2004.836239

where $\mathbf{x}^T$ represents the transpose of $\mathbf{x}$, $\mathbf{x}_i^j$ the jth sample of the ith class sample set, $\mathbf{u}_i$ the mean of the ith class sample set, and $\mathbf{u}$ the mean of all samples, where

$$\mathbf{u}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{x}_i^j, \quad \mathbf{u} = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i}\mathbf{x}_i^j.$$

According to [2] and [15], the FSODVs are generated using the following steps: the first one, denoted by $\boldsymbol{\omega}_1$, is the unit eigenvector that maximizes $J_F(\boldsymbol{\omega})$, where $J_F(\boldsymbol{\omega})$ is the Fisher criterion function defined as

$$J_F(\boldsymbol{\omega}) = \frac{\boldsymbol{\omega}^T\mathbf{S_B}\boldsymbol{\omega}}{\boldsymbol{\omega}^T\mathbf{S_W}\boldsymbol{\omega}}.$$

The $(i+1)$th one, denoted by $\boldsymbol{\omega}_{i+1}$, is the unit eigenvector that maximizes $J_F(\boldsymbol{\omega})$ under the following orthogonal constraints:

$$\boldsymbol{\omega}_{i+1}^T\boldsymbol{\omega}_j = 0, \quad (j = 1, 2, \ldots, i). \tag{2}$$

The FSODVs are obtained by repeating the above steps.

To extend the linear FSODV to the nonlinear case, let $\mathbf{X}$ be mapped into a feature space $\mathbf{F}$ through a nonlinear mapping function Φ

$$\Phi : \mathbf{X} \to \mathbf{F}, \quad \mathbf{x} \to \Phi(\mathbf{x}).$$

Denote the between-class scatter matrix $\mathbf{S_B}$, the within-class scatter matrix $\mathbf{S_W}$ and the total-class scatter matrix $\mathbf{S_T}$ in the feature space $\mathbf{F}$ by $\mathbf{S_B^\Phi}, \mathbf{S_W^\Phi}$, and $\mathbf{S_T^\Phi}$, respectively. Then we obtain

$$\begin{aligned}
\mathbf{S_B^\Phi} &= \sum_{i=1}^{c} N_i \left(\mathbf{u}_i^\Phi - \mathbf{u}^\Phi\right)\left(\mathbf{u}_i^\Phi - \mathbf{u}^\Phi\right)^T \\
\mathbf{S_W^\Phi} &= \sum_{i=1}^{c}\sum_{j=1}^{N_i}\left(\Phi\left(\mathbf{x}_i^j\right) - \mathbf{u}_i^\Phi\right)\left(\Phi\left(\mathbf{x}_i^j\right) - \mathbf{u}_i^\Phi\right)^T \\
\mathbf{S_T^\Phi} &= \sum_{i=1}^{c}\sum_{j=1}^{N_i}\left(\Phi\left(\mathbf{x}_i^j\right) - \mathbf{u}^\Phi\right)\left(\Phi\left(\mathbf{x}_i^j\right) - \mathbf{u}^\Phi\right)^T
\end{aligned} \tag{3}$$

where $\mathbf{u}_i^\Phi = (1/(N_i)\sum_{j=1}^{N_i}\Phi(\mathbf{x}_i^j)$ and $\mathbf{u}^\Phi = (1/N)\sum_{i=1}^{c}\sum_{j=1}^{N_i}\Phi(\mathbf{x}_i^j)$ are the mean of the ith class sample set and the mean of all samples in the feature space $\mathbf{F}$, respectively, and the Fisher discriminant criterion $J_F(\boldsymbol{\omega})$ in the feature space $\mathbf{F}$ can be rewritten as

$$J_F^\Phi(\boldsymbol{\omega}) = \frac{\boldsymbol{\omega}^T\mathbf{S_B^\Phi}\boldsymbol{\omega}}{\boldsymbol{\omega}^T\mathbf{S_W^\Phi}\boldsymbol{\omega}}. \tag{4}$$

Therefore, we obtain that the optimal discriminant vectors of KFSODV can be generated by repeating the following steps.

Step 1) The first optimal discriminant vector $\boldsymbol{\omega}_1^\Phi$ is the unit eigenvector that maximizes $J_F^\Phi(\boldsymbol{\omega})$ in $\mathbf{F}$;

Step 2) Suppose that $\boldsymbol{\omega}_1^\Phi, \boldsymbol{\omega}_2^\Phi, \ldots, \boldsymbol{\omega}_i^\Phi (i \geq 1)$ are the first i optimal discriminant vectors. Then the $(i+1)$th vector $\boldsymbol{\omega}_{i+1}^\Phi$ is the unit eigenvector that maximizes $J_F^\Phi(\boldsymbol{\omega})$ under the following orthogonal constraints:

$$\left(\boldsymbol{\omega}_{i+1}^\Phi\right)^T\boldsymbol{\omega}_j^\Phi = 0, \quad (j = 1, 2, \ldots, i). \tag{5}$$

It was shown [8] that the most discriminant vectors of LDA lie in the null space of the within-class scatter matrix in the case of the SSS problem. However, the denominator in the expression of $J_F^\Phi(\boldsymbol{\omega})$ may equal to zero when $\boldsymbol{\omega}$ lies in the null space of $\mathbf{S_W^\Phi}$. To overcome this problem, one way commonly used in previous literatures is to modify the discriminant criterion by using the total-class scatter matrix to replace the within-class scatter matrix in the Fisher criterion function [5], [9]. The other way avoiding this problem is to add a regularization matrix to the within-class scatter matrix such that the modified within-class matrix becomes invertible [16]. However, our further study shows that both approaches will face the degenerate eigenvalue problem (i.e., several eigenvectors share the same eigenvalue), which make the solutions not optimal in terms of discriminant ability [21], [22]. Different from the previous approaches, we present a new technique in this paper, which can directly overcome the SSS problem and at the same time avoid the degenerate eigenvalue perturbation problem.

B. Related Theorems

Let $\mathbf{S_T}(\mathbf{0})$ denote the null space of $\mathbf{S_T}$, $\overline{\mathbf{S_T}(\mathbf{0})}$ the orthogonal complement of $\mathbf{S_T}(\mathbf{0})$, and $\mathbf{I}$ the identity matrix. Then we have the following theorems.

Theorem 1: [5] Let $\mathbf{A}$ be a positive–semidefinite matrix. Then $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ iff $\mathbf{A}\mathbf{x} = \mathbf{0}$.

Theorem 2: [5] Let $\boldsymbol{\omega}_1$ be the first discriminant vector of FSODV. Then $\boldsymbol{\omega}_1$ can be chosen from $\overline{\mathbf{S_T}(\mathbf{0})}$.

Theorem 3: [17], [6] Let $\varphi_1, \ldots, \varphi_r$ be the first r discriminant vectors of FSODV. Then the $(r+1)$th vector φ_{r+1} is the eigenvector corresponding to the largest eigenvalue of the following eigenquation:

$$\mathbf{P}\mathbf{S_B}\varphi = \lambda\mathbf{S_W}\varphi$$

where

$$\begin{aligned}
\mathbf{P} &= \mathbf{I} - \mathbf{D}^T\left(\mathbf{D}\mathbf{S_W^{-1}}\mathbf{D}^T\right)^{-1}\mathbf{D}\mathbf{S_W^{-1}} \\
\mathbf{D} &= [\varphi_1 \quad \varphi_2 \quad \cdots \quad \varphi_r]^T.
\end{aligned}$$

We extended Theorem 2 and 3 to be more general forms and give the results as Theorem 4 and Theorem 5, respectively.

Theorem 4: Suppose that $\mathbf{S_T}$ is singular and let $\text{rank}(\overline{\mathbf{S_T}(\mathbf{0})}) = r$. Then all the discriminant vectors $\boldsymbol{\omega}_i (i = 1, 2, \ldots, r)$ of FSODV can be chosen from $\overline{\mathbf{S_T}(\mathbf{0})}$.

The Proof of Theorem 4 is given in Appendix I.

Theorem 5: Suppose that $\mathbf{B}$ and $\mathbf{R}$ are positive–semidefinite matrices, $\mathbf{V}$ is a positive matrix, and the discriminant criterion function is defined as

$$F(\varphi) = \frac{\varphi^T\mathbf{B}\varphi}{\varphi^T\mathbf{V}\varphi}. \tag{6}$$

Let φ_1 be the first unit discriminant eigenvector that maximizes $F(\varphi)$, and φ_{r+1} be the $(r+1)$th $(r \geq 1)$ unit discriminant eigenvector that maximizes $F(\varphi)$ under the following constraints:

$$\varphi_{r+1}^T\mathbf{R}\varphi_i = 0, \quad (i = 1, 2, \ldots, r). \tag{7}$$

Then φ_{r+1} is the unit eigenvector corresponding to the largest eigenvalue of the following eigenequation:

$$\mathbf{PB}\varphi = \lambda \mathbf{V}\varphi \tag{8}$$

where

$$\mathbf{P} = \mathbf{I} - \mathbf{R}\mathbf{D}^T(\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{R}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{R}\mathbf{V}^{-1} \tag{9}$$

$$\mathbf{D} = [\varphi_1 \quad \varphi_2 \quad \cdots \quad \varphi_r]^T. \tag{10}$$

The Proof of Theorem 5 is given in Appendix II.

III. Formulation of Kernel FSODV in Feature Space

Let

$$\Phi(\mathbf{X}) = \left[\Phi\left(\mathbf{x}_1^1\right)\cdots\Phi\left(\mathbf{x}_1^{N_1}\right)\cdots\Phi\left(\mathbf{x}_c^1\right)\cdots\Phi\left(\mathbf{x}_c^{N_c}\right)\right] \tag{11}$$

Then the expressions in (3) can be rewritten as

$$\mathbf{S}_\mathbf{B}^\Phi = \Phi(\mathbf{X})(\mathbf{L}-\mathbf{M})(\mathbf{L}-\mathbf{M})^T(\Phi(\mathbf{X}))^T \tag{12}$$

$$\mathbf{S}_\mathbf{W}^\Phi = \Phi(\mathbf{X})(\mathbf{I}-\mathbf{L})(\mathbf{I}-\mathbf{L})^T(\Phi(\mathbf{X}))^T \tag{13}$$

$$\mathbf{S}_\mathbf{T}^\Phi = \Phi(\mathbf{X})(\mathbf{I}-\mathbf{M})(\mathbf{I}-\mathbf{M})^T(\Phi(\mathbf{X}))^T \tag{14}$$

where $\mathbf{L} = (\mathbf{L}_l)_{l=1,\ldots,c}$ is a $N \times N$ block diagonal matrix, $\mathbf{L}_l$ is a $N_l \times N_l$ matrix with all terms equal to $1/N_l$, $\mathbf{I}$ is the $N \times N$ identity matrix, and $\mathbf{M} = (m_{ij})_{i=1,\ldots,N;j=1,\ldots,N}$ is a $N \times N$ matrix with all terms equal to $1/N$.

Assume that there exists a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ in $\mathbf{F}$, such that

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)\rangle = (\Phi(\mathbf{x}_i))^T\Phi(\mathbf{x}_j) \tag{15}$$

where $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)\rangle$ represents the dot product of $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$. For any two arbitrary classes p and q, we can express this kernel function by

$$(k_{ij})_{pq} = \langle \Phi\left(\mathbf{x}_p^i\right), \Phi\left(\mathbf{x}_q^j\right)\rangle = \left(\Phi\left(\mathbf{x}_p^i\right)\right)^T \Phi\left(\mathbf{x}_q^j\right). \tag{16}$$

Let $\mathbf{K}_{pq}$ be a $N_p \times N_q$ matrix defined by

$$\mathbf{K}_{pq} = (k_{ij})_{pq} \quad (i = 1, \ldots, N_p; j = 1, \ldots, N_q).$$

Let $\mathbf{K}$ be a $N \times N$ matrix defined by

$$\mathbf{K} = (\mathbf{K}_{pq})_{p=1,\ldots,c;q=1,\ldots,c}. \tag{17}$$

Then from (11), (16), and (17), we get

$$\mathbf{K} = (\Phi(\mathbf{X}))^T\Phi(\mathbf{X}). \tag{18}$$

Denote the null space of $\mathbf{S}_\mathbf{T}^\Phi$ and $\mathbf{S}_\mathbf{W}^\Phi$ by $\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})$ and $\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$, respectively. Let $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})}$ and $\overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})}$ represent the orthogonal complement of $\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})$ and $\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$, respectively. From Theorem 4, we obtain that the discriminant vectors of KFSODV can be chosen from $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})}$. Besides, from the analysis in Section II, we know that the most discriminant vectors of KFSODV lie in $\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$ as to the case of the SSS problem. According to the above analysis we, thus, proposed to use the following steps to solve the discriminant vectors of KFSODV:

Step 1) resolve the basis vectors of the subspace $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})}$;
Step 2) resolve KFSODV in the subspace $\overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})}$;
Step 3) resolve KFSODV in the subspace $\overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})}$.

A. Basis of Subspace $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})}$, $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})} \cap \mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$, and $\overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})}$

An effective way to get the basis of $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})}$ is to perform the eigenvector decomposition using the KPCA method [10]. KPCA implements this by finding the eigenvectors $\boldsymbol{\omega}$ and the corresponding eigenvalues $\lambda \geq 0$, solutions of the eigenequation $\lambda\boldsymbol{\omega} = \mathbf{S}_\mathbf{T}^\Phi\boldsymbol{\omega}$. From [10], we know that the eigenvectors $\boldsymbol{\omega}$ can be written as follows:

$$\boldsymbol{\omega} = \sum_{p=1}^{c}\sum_{q=1}^{N_p}\alpha_{pq}\left(\Phi\left(\mathbf{x}_p^q\right) - \mathbf{u}^\Phi\right) = \Phi(\mathbf{X})(\mathbf{I}-\mathbf{M})\boldsymbol{\alpha} \tag{19}$$

where $\boldsymbol{\alpha} = [\alpha_{11}\ \alpha_{12}\ \cdots\ \alpha_{cN_c}]^T$.

KPCA then turns to resolve the following eigenequation:

$$\lambda\boldsymbol{\alpha} = \tilde{\mathbf{K}}\boldsymbol{\alpha} \tag{20}$$

where

$$\tilde{\mathbf{K}} = (\mathbf{I}-\mathbf{M})^T\mathbf{K}(\mathbf{I}-\mathbf{M}). \tag{21}$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ denote the eigenvalues of $\tilde{\mathbf{K}}$, and $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$ the corresponding complete set of eigenvectors, with λ_r being the last nonzero eigenvalue. Then we obtain that

$$\boldsymbol{\omega}_i = \Phi(\mathbf{X})(\mathbf{I}-\mathbf{M})\boldsymbol{\alpha}_i (i = 1, \ldots, r) \tag{22}$$

are the basis vectors of $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})}$. The coefficients $\boldsymbol{\alpha}_i (i = 1, \ldots, r)$ are divided by $\sqrt{\lambda_i}$, respectively, in order to get the normalized vectors $\boldsymbol{\omega}_i (i = 1, \ldots, r)$ in $\mathbf{F}$. That is

$$\boldsymbol{\omega}_i^T\boldsymbol{\omega}_i = 1. \tag{23}$$

Let

$$\mathbf{Q} = [\boldsymbol{\omega}_1 \quad \boldsymbol{\omega}_2 \quad \cdots \quad \boldsymbol{\omega}_r], \quad \mathbf{H} = [\boldsymbol{\alpha}_1 \quad \boldsymbol{\alpha}_2 \quad \cdots \quad \boldsymbol{\alpha}_r.] \tag{24}$$

From (19) and (24), the matrix $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \Phi(\mathbf{X})(\mathbf{I}-\mathbf{M})\mathbf{H}. \tag{25}$$

Because $\boldsymbol{\omega}_i (i = 1, \ldots, r)$ are orthogonal each other, we get

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{H}^T(\mathbf{I}-\mathbf{M})^T\mathbf{K}(\mathbf{I}-\mathbf{M})\mathbf{H} = \mathbf{I}_r \tag{26}$$

where $\mathbf{I}_r$ is the $r \times r$ identity matrix.

In what follows, we divide the subspace $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})}$ into two orthogonal subspace, i.e., $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})} \cap \mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$ and $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})} \cap \overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})}$, and then resolve the basis vectors of the two subspace, respectively. Noting that $\boldsymbol{\omega}_i (i = 1, \ldots, r)$ form a basis of $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})}$, thus, for any arbitrary vector $\boldsymbol{\xi} \in (\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})} \cap \mathbf{S}_\mathbf{W}^\Phi(\mathbf{0}))$, there exist coefficients $\beta_i (i = 1, \ldots, r)$ such that $\boldsymbol{\xi} = \sum_{i=1}^r \boldsymbol{\omega}_i\beta_i$ and

$$\boldsymbol{\xi}^T\mathbf{S}_W^\Phi\boldsymbol{\xi} = 0. \tag{27}$$

Let $\boldsymbol{\beta} = [\beta_1 \ \cdots \ \beta_r]^T$. From (25), we obtain

$$\boldsymbol{\xi} = \mathbf{Q}\boldsymbol{\beta} = \Phi(\mathbf{X})(\mathbf{I} - \mathbf{M})\mathbf{H}\boldsymbol{\beta}. \tag{28}$$

From (13), (18), (27), and (28), we get

$$\boldsymbol{\beta}^T \mathbf{W} \boldsymbol{\beta} = 0 \tag{29}$$

where

$$\mathbf{W} = \mathbf{H}^T(\mathbf{I} - \mathbf{M})^T\mathbf{K}(\mathbf{I} - \mathbf{L})(\mathbf{I} - \mathbf{L})^T\mathbf{K}(\mathbf{I} - \mathbf{M})\mathbf{H}. \tag{30}$$

From Theorem 1 and (29), we get

$$\mathbf{W}\boldsymbol{\beta} = \mathbf{0}. \tag{31}$$

Equation (31) means that $\boldsymbol{\beta}$ is the eigenvector corresponding to the zero eigenvalue of the matrix $\mathbf{W}$.

Let $\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_r$ be the eigenvalues of $\mathbf{W}$, and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r$ the corresponding complete set of orthogonal eigenvectors, with $\tilde{\lambda}_s$ (without loss of generality, we assume that $s \geq 1$) being the last zero eigenvalue. Denote $\boldsymbol{\xi}_i = \Phi(\mathbf{X})(\mathbf{I} - \mathbf{M})\mathbf{H}\boldsymbol{\beta}_i (i = 1, \ldots, r)$. Then the vectors $\boldsymbol{\xi}_i (i = 1, \ldots, s)$ form a basis of $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})} \cap \mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$, and the remainder vectors $\boldsymbol{\xi}_i (i = s+1, \ldots, r)$ form a basis of $\overline{\mathbf{S}_\mathbf{T}^\Phi(\mathbf{0})} \cap \overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})} (= \overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})})$. Let

$$\mathbf{U} = [\boldsymbol{\beta}_1 \quad \cdots \quad \boldsymbol{\beta}_s], \quad \mathbf{V} = [\boldsymbol{\beta}_{s+1} \quad \cdots \quad \boldsymbol{\beta}_r] \tag{32}$$

then we have

$$[\boldsymbol{\xi}_1 \quad \cdots \quad \boldsymbol{\xi}_s] = \Phi(\mathbf{X})(\mathbf{I} - \mathbf{M})\mathbf{H}\mathbf{U} \tag{33}$$

$$[\boldsymbol{\xi}_{s+1} \quad \cdots \quad \boldsymbol{\xi}_r] = \Phi(\mathbf{X})(\mathbf{I} - \mathbf{M})\mathbf{H}\mathbf{V}. \tag{34}$$

Furthermore, consider that $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r$ are orthogonal each other, thus, we have

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_s \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_{r-s} \tag{35}$$

where $\mathbf{I}_s$ is the $s \times s$ identity matrix and $\mathbf{I}_{r-s}$ is the $(r-s) \times (r-s)$ identity matrix.

B. KFSODVs in Subspace $\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$

According to discriminant analysis [7], [14], we obtain that the optimal discriminant vectors should maximize the between-class scatter and simultaneously minimize the within-class scatter. However, as to the extreme case where the optimal vectors lie in $\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$, new criterion should be defined to replace the previous discriminant criterion to avoid the degenerate eigenvalue problem [21], [22]. Such a criterion can be defined as

$$J_1^\Phi(\boldsymbol{\omega}) = \boldsymbol{\omega}^T \mathbf{S}_\mathbf{B}^\Phi \boldsymbol{\omega}. \tag{36}$$

Suppose that $\boldsymbol{\omega}^*$ is a discriminant vector of KFSODV in $\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})$. Then there exist coefficients $\gamma_i (i = 1, \ldots, s)$ such that $\boldsymbol{\omega}^* = \sum_{i=1}^s \boldsymbol{\xi}_i \gamma_i$. From (33), we obtain that $\boldsymbol{\omega}^*$ can be rewritten as

$$\boldsymbol{\omega}^* = \Phi(\mathbf{X})(\mathbf{I} - \mathbf{M})\mathbf{H}\mathbf{U}\boldsymbol{\gamma}. \tag{37}$$

where $\boldsymbol{\gamma}$ denotes the column vector with entries $\gamma_1, \ldots, \gamma_s$.

From (12), (18), and (37), we get

$$J_1^\Phi(\boldsymbol{\omega}^*) = (\boldsymbol{\omega}^*)^T \mathbf{S}_\mathbf{B}^\Phi \boldsymbol{\omega}^* = \boldsymbol{\gamma}^T \mathbf{B} \boldsymbol{\gamma}$$

where

$$\mathbf{B} = \mathbf{U}^T\mathbf{H}^T(\mathbf{I} - \mathbf{M})^T\mathbf{K}(\mathbf{L} - \mathbf{M})(\mathbf{L} - \mathbf{M})^T\mathbf{K}(\mathbf{I} - \mathbf{M})\mathbf{H}\mathbf{U}. \tag{38}$$

Let $\tilde{\tilde{\lambda}}_1 \geq \tilde{\tilde{\lambda}}_2 \geq \cdots \geq \tilde{\tilde{\lambda}}_s$ be the eigenvalues of $\mathbf{B}$, and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s$ the corresponding complete set of unit eigenvectors. Denote

$$\boldsymbol{\omega}_i^* = \Phi(\mathbf{X})(\mathbf{I} - \mathbf{M})\mathbf{H}\mathbf{U}\boldsymbol{\gamma}_i, \quad (i = 1, \ldots, s) \tag{39}$$

then we obtain

$$J_1^\Phi(\boldsymbol{\omega}_i^*) = \boldsymbol{\gamma}_i^T \mathbf{B} \boldsymbol{\gamma}_i = \tilde{\tilde{\lambda}}_i$$
$$J_1^\Phi(\boldsymbol{\omega}_1^*) \geq J_1^\Phi(\boldsymbol{\omega}_2^*) \geq \cdots \geq J_1^\Phi(\boldsymbol{\omega}_s^*).$$

Besides, from (26) and (35), we get

$$\begin{aligned}(\boldsymbol{\omega}_i^*)^T\boldsymbol{\omega}_j^* &= \boldsymbol{\gamma}_i^T\mathbf{U}^T\mathbf{H}^T(\mathbf{I} - \mathbf{M})^T\mathbf{K}(\mathbf{I} - \mathbf{M})\mathbf{H}\mathbf{U}\boldsymbol{\gamma}_j \\ &= \boldsymbol{\gamma}_i^T\boldsymbol{\gamma}_j = \delta_{ij}\end{aligned} \tag{40}$$

where

$$\delta_{ij} = \begin{cases} 1, & (i = j) \\ 0, & (i \neq j) \end{cases}.$$

Equation (40) also means that $\boldsymbol{\omega}_i^*$ is a unit vector in the feature space $\mathbf{F}$. According to the definition of KFSODV, it is reasonable to choose $\boldsymbol{\omega}_1^*$ as the first vector of KFSODV, and $\boldsymbol{\omega}_i^* (i \leq s)$ as the ith one. In this way, we can obtain the first s optimal discriminant vectors of KFSODV.

The projection of a test point $\mathbf{z}$ onto the vectors $\boldsymbol{\omega}_i^*$ can be calculated by

$$(\boldsymbol{\omega}_i^*)^T\Phi(\boldsymbol{z}) = \boldsymbol{\gamma}_i^T\mathbf{U}^T\mathbf{H}^T(\mathbf{I} - \mathbf{M})\boldsymbol{\kappa} \tag{41}$$

where $\boldsymbol{\kappa}$ is a $N \times 1$ vector defined as

$$\boldsymbol{\kappa} = \left[\kappa_1^1 \quad \cdots \quad \kappa_1^{N_1} \quad \cdots \quad \kappa_c^1 \quad \cdots \quad \kappa_c^{N_c}\right]^T \tag{42}$$

where $\kappa_i^j = k(\mathbf{x}_i^j, \mathbf{z})$.

C. KFSODVs in Subspace $\overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})}$

Similar to the analysis in Section III-B, for any optimal discriminant vector φ^* of KFSODV in $\overline{\mathbf{S}_\mathbf{W}^\Phi(\mathbf{0})}$, there exist

coefficients $\gamma_i (i = s+1, \ldots, r)$ such that $\varphi^* = \sum_{i=s+1}^{r} \boldsymbol{\xi}_i \gamma_i$. From (34), we obtain

$$\varphi^* = \Phi(\mathbf{X})(\mathbf{I} - \mathbf{M})\mathbf{H}\mathbf{V}\boldsymbol{\gamma}^* \tag{43}$$

where $\boldsymbol{\gamma}^*$ denotes the column vector with entries $\gamma_{s+1}, \ldots, \gamma_r$.

Substituting $\varphi = \Phi(\mathbf{X})(\mathbf{I}-\mathbf{M})\mathbf{H}\mathbf{V}\boldsymbol{\gamma}$ into the Fisher's criterion $J_F^{\Phi}(\varphi)$, we get

$$J_F^{\Phi}(\varphi) = \frac{\varphi^T \mathbf{S}_{\mathbf{B}}^{\Phi} \varphi}{\varphi^T \mathbf{S}_{\mathbf{W}}^{\Phi} \varphi} = \frac{\boldsymbol{\gamma}^T \mathbf{B}_1 \boldsymbol{\gamma}}{\boldsymbol{\gamma}^T \mathbf{W}_1 \boldsymbol{\gamma}} = \tilde{J}_F^{\Phi}(\boldsymbol{\gamma}) \tag{44}$$

where

$$\mathbf{B}_1 = \mathbf{V}^T \mathbf{H}^T (\mathbf{I}-\mathbf{M})^T \mathbf{K} (\mathbf{L}-\mathbf{M})(\mathbf{L}-\mathbf{M})^T \mathbf{K} (\mathbf{I}-\mathbf{M})\mathbf{H}\mathbf{V} \tag{45}$$

$$\mathbf{W}_1 = \mathbf{V}^T \mathbf{H}^T (\mathbf{I}-\mathbf{M})^T \mathbf{K} (\mathbf{I}-\mathbf{L})(\mathbf{I}-\mathbf{L})^T \mathbf{K} (\mathbf{I}-\mathbf{M})\mathbf{H}\mathbf{V}. \tag{46}$$

Let $\varphi_1^* = \Phi(\mathbf{X})(\underline{\mathbf{I}-\mathbf{M}})\mathbf{H}\mathbf{V}\boldsymbol{\gamma}_1^*$ be the first discriminant vector of KFSODV in $\overline{\mathbf{S}_{\mathbf{W}}^{\Phi}(\mathbf{0})}$. Then $\boldsymbol{\gamma}_1^*$ is the unit eigenvector that maximizes $\tilde{J}_F^{\Phi}(\boldsymbol{\gamma})$, or equivalently, $\boldsymbol{\gamma}_1^*$ is the unit eigenvector corresponding to the largest eigenvalue of the eigenequation

$$\mathbf{B}_1 \boldsymbol{\gamma} = \lambda \mathbf{W}_1 \boldsymbol{\gamma}. \tag{47}$$

Suppose that we have obtained the first t optimal discriminant vectors $\varphi_i^* = \Phi(\mathbf{X})(\mathbf{I}-\mathbf{M})\mathbf{H}\mathbf{V}\boldsymbol{\gamma}_i^* (i = 1, \ldots, t)$, then to find the $(t+1)$th vector $\varphi_{t+1}^* = \Phi(\mathbf{X})(\mathbf{I}-\mathbf{M})\mathbf{H}\mathbf{V}\boldsymbol{\gamma}_{t+1}^*$ is equivalent to finding the unit eigenvector $\boldsymbol{\gamma}_{t+1}^*$ that maximizes $\tilde{J}_F^{\Phi}(\boldsymbol{\gamma})$ under the orthogonal constraints

$$0 = (\varphi_{t+1}^*)^T \varphi_i^* = (\boldsymbol{\gamma}_{t+1}^*)^T \mathbf{R} \boldsymbol{\gamma}_i^*, \quad (i = 1, 2, \ldots, t) \tag{48}$$

where

$$\mathbf{R} = \mathbf{V}^T \mathbf{H}^T (\mathbf{I}-\mathbf{M})^T \mathbf{K} (\mathbf{I}-\mathbf{M})\mathbf{H}\mathbf{V}. \tag{49}$$

From Theorem 5, we obtain that $\boldsymbol{\gamma}_{t+1}^*$ is the eigenvector corresponding to the largest eigenvalue of the following eigenequation:

$$\mathbf{P}\mathbf{B}_1 \boldsymbol{\gamma} = \lambda \mathbf{W}_1 \boldsymbol{\gamma} \tag{50}$$

where

$$\mathbf{P} = \mathbf{I} - \mathbf{R}\mathbf{D}^T \left(\mathbf{D}\mathbf{R}\mathbf{W}_1^{-1}\mathbf{R}\mathbf{D}^T \right)^{-1} \mathbf{D}\mathbf{R}\mathbf{W}_1^{-1} \tag{51}$$

$$\mathbf{D} = [\boldsymbol{\gamma}_1^* \quad \boldsymbol{\gamma}_2^* \quad \cdots \quad \boldsymbol{\gamma}_t^*]^T. \tag{52}$$

By repeating the above steps, we can obtain the direction of each discriminant vector of KFSODV in $\overline{\mathbf{S}_{\mathbf{W}}^{\Phi}(\mathbf{0})}$. Moreover, from (26) and (35), we have

$$\begin{aligned}(\varphi_i^*)^T \varphi_i^* &= (\boldsymbol{\gamma}_i^*)^T \mathbf{V}^T \mathbf{H}^T (\mathbf{I}-\mathbf{M})^T \mathbf{K} (\mathbf{I}-\mathbf{M})\mathbf{H}\mathbf{V}\boldsymbol{\gamma}_i^* \\ &= (\boldsymbol{\gamma}_i^*)^T \boldsymbol{\gamma}_i^* = 1.\end{aligned} \tag{53}$$

Equation (53) means that φ_i^* is a unit vector. Thus, φ_i^* is the discriminant vector of KFSODV in $\overline{\mathbf{S}_{\mathbf{W}}^{\Phi}(\mathbf{0})}$.

The projection of a test point $\mathbf{z}$ onto the discriminant vectors φ_i^* can be calculated by

$$(\varphi_i^*)^T \Phi(\mathbf{z}) = (\boldsymbol{\gamma}_i^*)^T \mathbf{V}^T \mathbf{H}^T (\mathbf{I}-\mathbf{M})\boldsymbol{\kappa}. \tag{54}$$

The procedure of computing KFSODV can be summarized in the following steps:

Step 1) compute the kernel matrices $\mathbf{K}$ and $\tilde{\mathbf{K}}$ using (18) and (21);
Step 2) decompose $\tilde{\mathbf{K}}$ using eigenvectors decomposition and normalize the eigenvectors;
Step 3) compute the matrices $\mathbf{H}$ and $\mathbf{W}$ using (24) and (30);
Step 4) decompose $\mathbf{W}$ using eigenvectors decomposition. Let s be the number of the eigenvectors corresponding to the zero eigenvalues;
If $s \geq 1$, then
a) compute the matrices $\mathbf{U}$ and $\mathbf{B}$ using (32) and (38);
b) decompose $\mathbf{B}$ using eigenvectors decomposition to get the eigenvectors $\boldsymbol{\gamma}_i$;
c) compute projection of a test point $\mathbf{z}$ onto the discriminant vectors $\boldsymbol{\omega}_i^*$ using (41).
Step 5) compute the matrices $\mathbf{V}, \mathbf{B}_1, \mathbf{W}_1$, and $\mathbf{R}$ using (32), (45), (46), and (49);
Step 6) compute eigenvectors $\boldsymbol{\gamma}_i^* (i = 1, 2, \ldots)$ using the systems (47) and (50);
Step 7) compute projections of a test point $\mathbf{z}$ onto the discriminant vector φ_i^* using (54).

IV. Experiments

In this section, we will use simulated and real data sets to test the performance of the proposed method against the commonly used kernel-based learning algorithms: KPCA, GDA, and kernel direct discriminant analysis (KDDA) recently proposed by Lu *et al.* [18]. The polynomial and gaussian kernels defined as follows are used to compute the elements of the matrix $\mathbf{K}$: $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ in these experiments:

Polynomial kernel: $k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle)^d$, where d is the polynomial degree;
Gaussian kernel: $k(\mathbf{x}, \mathbf{y}) = \exp(-||\mathbf{x}-\mathbf{y}||^2/\sigma)$, where σ is chosen empirically in the experiments of this paper.

A. Comparison to KPCA Using Simulated Data

Two data sets with parabolic shapes vertically and horizontally mirrored in this experiment were generated by the function $y = x^2 + 0.5 + \zeta$, where the x values have a uniform distribution in $[-1, 1]$, and ζ is a uniformly distributed random number on the interval $[0, 0.01]$. We used 100 data points uniformly divided for each parabolic shape to obtain 200 points as the training samples and then computed the discriminant vectors using KFSODV and KPCA, respectively, based on the gaussian kernel with parameter $\sigma = 1.0$. By projecting the test data onto the discriminant vectors, we can visually compare the discrimination performance between the two methods. Fig. 1 illustrates the results of the first four features extracted by KPCA and KFSODV, respectively. Depicted in the figures are the feature values (indicated by gray level) and contour lines of identical feature values. From Fig. 1 we can clearly see that the features extracted by KFSODV contain better discriminant information than those by KPCA. Especially, the first three features extracted by KFSODV can discriminate the two sample sets in a nearly optimal way, whereas those by KPCA do not separate them well.

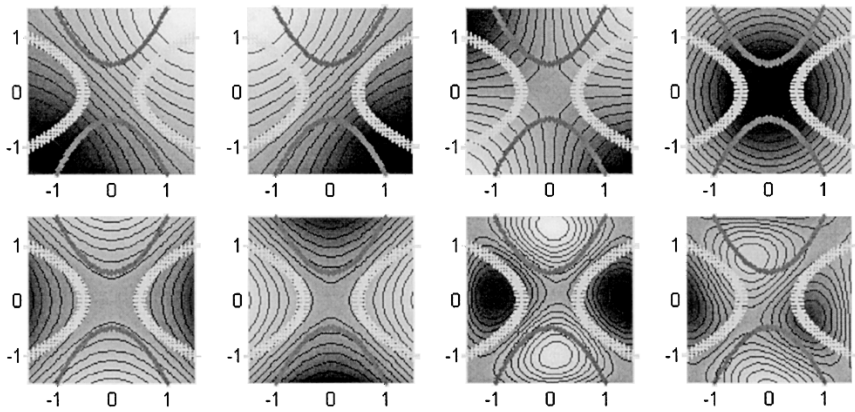

Fig. 1. First four features (ordered from left to right) extracted by KPCA (top) and KFSODV (bottom).

B. Comparison to Other Kernel-Based Methods Using Fisher's Iris Data

The well-known Fisher's Iris flower data set originally published in [4] was used in this experiment. The data set contains four measurements measured in millimeters on 50 Iris specimens from each of three species: Iris setosa, Iris versicolor, and Iris virginica [9]. It was shown in [9] that the first one was linearly separable from the latter two, whereas the latter two were not linearly separable from each other. Thus, we only choose the latter two species, that is, the Iris versicolor and Iris virginica, as the experiment data sets in this experiment to compare the discriminant ability among the commonly used kernel-based learning methods.

Fig. 2 shows the projections of the Iris versicolor data set and Iris virginica data set onto the first two eigenvectors computed by KPCA and KFSODV associated with polynomial kernel with $d = 2$. From Fig. 2, we can see that the projections onto the eigenvectors of KPCA have much larger variance than those onto the discriminant vectors of KFSODV. As for the discriminant information, however, the projections onto the discriminant vectors of KFSODV achieve better results than those onto the eigenvectors of KPCA.

To further compare the discriminant performance of KFSODV with other commonly used kernel-based learning algorithms, we adopted the "leave-one-out" strategy to perform classification problem, where the polynomial kernel with $d = 2$ and gaussian kernel with $\sigma = 10\,000$ were used in this experiment. In the "leave-one-out" method, one sample was excluded as the test sample and the remains (99 samples here) as the training samples to compute the projected vectors. This operation was repeated 100 trials to test all the 100 samples. Then the number of the misclassified samples was counted to obtain the estimate of the test error rate [7]. Fig. 3 shows the relationship plot between the test error rate and the number of features extracted by KPCA and KFSODV, respectively, whereas Table I shows the test error rates of the experiments.

From Fig. 3, we can see that 1) KFSODV achieves much lower test error rate than KPCA (1% versus 5%) and 2) the best result of KFSODV is obtained when the first two features are used for polynomial kernel. Furthermore, from Table I we can see that KFSODV achieves the best result $(=1\%)$ among all the four methods. Additionally, from Fig. 3 and Table I we also see that KFSODV achieves the same test error rate as GDA (2% for polynomial kernel and 6% for gaussian kernel) when only the first feature is used.

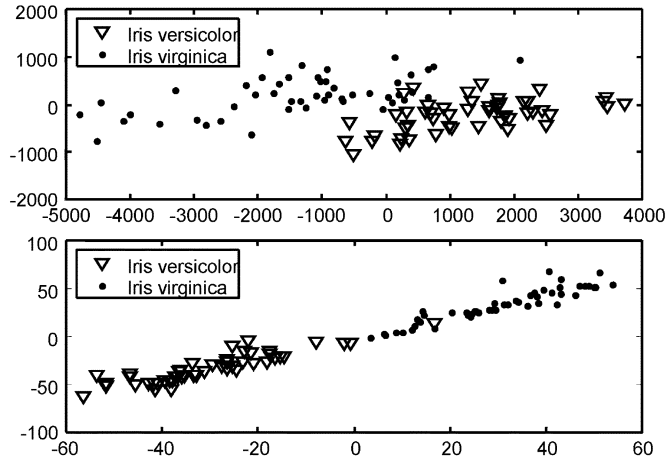

Fig. 2. Projections of Iris versicolor and Iris virginica onto the first two projected axes computed by KPCA (top) and KFSODV (bottom) with polynomial kernel $(d = 2)$.

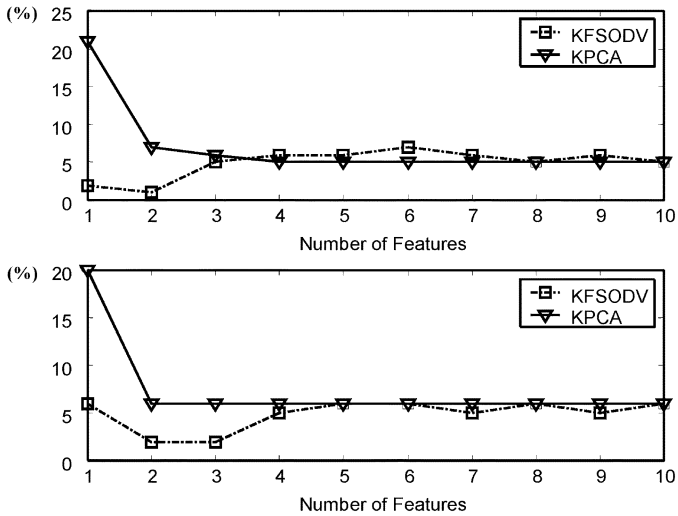

Fig. 3. Compare the test error rate between KPCA and KFSODV as the function of the number of features based on polynomial kernel with $d = 2$ (top) and Gaussian kernel with $\sigma = 10\,000$ (bottom).

TABLE I
COMPARISON OF VARIOUS SYSTEMS FOR TEST ERROR RATE IN IRIS VERSICOLOR DATA AND IRIS VIRGINICA DATA

Method	Reduced Space	Error Rate (%)
KDDA, $d = 2$	1	13.00 (13/100)
KDDA, $\sigma = 10000$	1	14.00 (14/100)
GDA, $d = 2$	1	2.00 (2/100)
GDA, $\sigma = 10000$	1	6.00 (6/100)
KPCA, $d = 2$	4	5.00 (5/100)
KPCA, $\sigma = 10000$	2	6.00 (6/100)
KFSODV, $d = 2$	2	**1.00** (1/100)
KFSODV, $\sigma = 10000$	2	2.00 (2/100)

C. Comparison to Other Kernel-Based Methods in Face Recognition

Face recognition is a very popular and efficient way to test the performance of feature extraction method. In this experiment, we test the KFSODV method on the Yale face database. The Yale face database contains 165 face images of 15 subjects that include variations in both facial expression and lighting condition [19]. The original face images were sized 243×320 pixels with a 256-level gray scale. Fig. 4 shows some images of one subject in the database. To reduce the computational complexity, we resized the face images to be 92×112 pixels, a commonly used image size in face recognition tasks [18]. Then downsampled them to be 23×28 pixels by using the wavelet transformation method [20] and normalized them using a linear function. After doing that, the mean and standard deviation of Kurtosis of the images were 1.86 and 0.32, respectively.

We adopted the "leave-one-out" strategy and used KFSODV, KPCA, KDDA, and GDA, respectively, to perform the face recognition. The polynomial kernel with $d = 2$ and gaussian kernel with $\sigma = 10\,000$ were used. Table II shows the experimental results. From Table II we see that the KFSODV method, when 30 features used, achieves the lowest test error rate $(=6.06\%)$ among the four methods.

V. DISCUSSION AND CONCLUSION

In this paper, the FSODV method has been extended from linear domain to a nonlinear domain via the kernel trick. Similar to the relationship between PCA and KPCA [10], many of the mathematical and statistical properties of FSODV may carry over to KFSODV. The two main points are 1) KFSODV is an orthogonal basis transformation method in the feature space and 2) the discriminant vectors of KFSODV carry more discriminant information than other orthogonal vectors in the feature space in the sense of the Fisher discriminant criterion. The latter point also means that the KFSODV transformation could be superior to KPCA in terms of the discriminant ability. In fact, the similar and competitive method to KFSODV may be the GDA method, which is another nonlinear feature extraction method based on the same Fisher's criterion in the feature space [6]. However, the performance of GDA may be limited due to the limitation of the number of the classes. This is because the number of the discriminant vectors computed by GDA generally depends on the number of the classes of the sample sets. By contrast,

Fig. 4. Face images of one subject from the Yale face database.

TABLE II
COMPARISON OF VARIOUS SYSTEMS FOR TEST ERROR RATE IN YALE FACE DATABASE

Method	Reduced Space	Error Rate (%)
KDDA, $d = 2$	14	16.36 (27/165)
KDDA, $\sigma = 10000$	14	16.97 (28/165)
GDA, $d = 2$	14	9.09 (15/165)
GDA, $\sigma = 10000$	14	8.48 (14/165)
KPCA, $d = 2$	30	23.03 (38/165)
KPCA, $\sigma = 10000$	30	22.42 (37/165)
KFSODV, $d = 2$	14	9.09 (15/165)
KFSODV, $\sigma = 10000$	30	**6.06** (10/165)

KFSODV can obtain more discriminant vectors than GDA and, therefore, may obtain better performance than GDA when the number of the classes is small.

Additionally, our experimental results also show that GDA achieves better performance than KDDA, which seems to contradict to the experimental results reported in [18]. One explanation for this could attribute to the fact that KDDA computes the discriminant vectors in the orthogonal complement of the null space of the between-class scatter matrix. However, as to the SSS problem, the most discriminant vectors lying in the null space of the within-class scatter matrix lie neither in the null space of the between-class scatter matrix nor in its orthogonal complement. Thus, it may fail to obtain these optimal discriminant vectors by using the KDDA algorithm.

APPENDIX I

Proof of Theorem 4:

Proof: Let $\boldsymbol{\omega}_1$ be the first discriminant vector of FSODV. From the Theorem 3, we obtain that $\boldsymbol{\omega}_1$ can be chosen from $\overline{\mathbf{S_T}(\mathbf{0})}$. Suppose that $\boldsymbol{\omega}_i, i = 1, 2, \ldots, r \geq 1$ are the first r vectors of FSODV, and $\boldsymbol{\omega}_i \in \overline{\mathbf{S_T}(\mathbf{0})}$. Let $\boldsymbol{\omega}_{r+1}$ be the $(r+1)$th discriminant vector of FSODV.

If $\boldsymbol{\omega}_{r+1} \notin \overline{\mathbf{S_T}(\mathbf{0})}$, let

$$\boldsymbol{\omega}_{r+1} = \boldsymbol{\omega}_{r+1}^{(1)} + \boldsymbol{\omega}_{r+1}^{(2)} \tag{A.1}$$

where $\boldsymbol{\omega}_{r+1}^{(1)} \in \overline{\mathbf{S_T}(\mathbf{0})}$ and $\boldsymbol{\omega}_{r+1}^{(2)} \in \mathbf{S_T}(\mathbf{0})$. That is, $\mathbf{S_T}\boldsymbol{\omega}_{r+1}^{(2)} = \mathbf{0}$.

From Theorem 1, we obtain that

$$\left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_T}\boldsymbol{\omega}_{r+1}^{(2)} = 0. \tag{A.2}$$

Consider that $\mathbf{S_T} = \mathbf{S_W} + \mathbf{S_B}$, thus, we have

$$\left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(2)} + \left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(2)} = 0. \tag{A.3}$$

Since $\mathbf{S_B}$ and $\mathbf{S_W}$ are positive–semidefinite matrices, we have

$$\left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(2)} \geq 0, \quad \left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(2)} \geq 0. \quad \text{(A.4)}$$

From (A.3) and (A.4), we obtain

$$\left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(2)} = 0, \quad \left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(2)} = 0. \quad \text{(A.5)}$$

From Theorem 1, (A.5) can be expressed as follows:

$$\mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(2)} = \mathbf{0}, \quad \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(2)} = \mathbf{0}. \quad \text{(A.6)}$$

Therefore, we have

$$\begin{aligned}\boldsymbol{\omega}_{r+1}^T \mathbf{S_B}\boldsymbol{\omega}_{r+1} &= \left(\boldsymbol{\omega}_{r+1}^{(1)} + \boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_B} \left(\boldsymbol{\omega}_{r+1}^{(1)} + \boldsymbol{\omega}_{r+1}^{(2)}\right) \\ &= \left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(1)} + \left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(2)} \\ &\quad + \left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(1)} + \left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(2)} \\ &= \left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(1)} \quad \text{(A.7)} \\ \boldsymbol{\omega}_{r+1}^T \mathbf{S_W}\boldsymbol{\omega}_{r+1} &= \left(\boldsymbol{\omega}_{r+1}^{(1)} + \boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_W} \left(\boldsymbol{\omega}_{r+1}^{(1)} + \boldsymbol{\omega}_{r+1}^{(2)}\right) \\ &= \left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(1)} + \left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(2)} \\ &\quad + \left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(1)} + \left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(2)} \\ &= \left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(1)}. \quad \text{(A.8)}\end{aligned}$$

From (A.7) and (A.8), we obtain

$$\begin{aligned}J(\boldsymbol{\omega}_{r+1}) &= \frac{\boldsymbol{\omega}_{r+1}^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}}{\boldsymbol{\omega}_{r+1}^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}} \\ &= \frac{\left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \mathbf{S_B}\boldsymbol{\omega}_{r+1}^{(1)}}{\left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \mathbf{S_W}\boldsymbol{\omega}_{r+1}^{(1)}} = J\left(\boldsymbol{\omega}_{r+1}^{(1)}\right). \quad \text{(A.9)}\end{aligned}$$

Besides, we have

$$\left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \boldsymbol{\omega}_i + \left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \boldsymbol{\omega}_i = \boldsymbol{\omega}_{r+1}^T \boldsymbol{\omega}_i = 0, \quad i \leq r. \quad \text{(A.10)}$$

Note that $\boldsymbol{\omega}_{r+1}^{(2)} \in \mathbf{S_T}(\mathbf{0})$ and $\boldsymbol{\omega}_i \in \overline{\mathbf{S_T}(\mathbf{0})} (i \leq r)$. Thus, we have

$$\left(\boldsymbol{\omega}_{r+1}^{(2)}\right)^T \boldsymbol{\omega}_i = 0, \quad i \leq r. \quad \text{(A.11)}$$

From (A.10) and (A.11), we obtain

$$\left(\boldsymbol{\omega}_{r+1}^{(1)}\right)^T \boldsymbol{\omega}_i = 0, \quad i \leq r. \quad \text{(A.12)}$$

From (A.9) and (A.12), we obtain that $\boldsymbol{\omega}_{r+1}^{(1)}$ (when normalized) is also the $(r+1)$th discriminant vector of FSODV.

According to the analysis above, we obtain that all the discriminant vectors of FSODV can be chosen from $\overline{\mathbf{S_T}(\mathbf{0})}$. □

APPENDIX II

Proof of Theorem 5:

Proof: Let $\tilde{\varphi}_{r+1}$ be the direction of φ_{r+1} and $\tilde{\varphi}_{r+1}$ satisfy the following constraint:

$$\tilde{\varphi}_{r+1}^T \mathbf{V} \tilde{\varphi}_{r+1} = 1. \quad \text{(B.1)}$$

Thus, from (7), we have

$$\tilde{\varphi}_{r+1}^T \mathbf{R} \varphi_i = 0, \quad i = 1, \ldots, r. \quad \text{(B.2)}$$

In order to compute $\tilde{\varphi}_{r+1}$, we use the method of Lagrange multipliers to transform the criterion (6) including all the constrains

$$\begin{aligned}L(\tilde{\varphi}_{r+1}) = \tilde{\varphi}_{r+1}^T \mathbf{B} \tilde{\varphi}_{r+1} - \lambda \left[\tilde{\varphi}_{r+1}^T \mathbf{V} \tilde{\varphi}_{r+1} - 1\right] \\ - \sum_{i=1}^{r} \mu_i \tilde{\varphi}_{r+1}^T \mathbf{R} \varphi_i \quad \text{(B.3)}\end{aligned}$$

where λ and $\mu_i, i = 1, 2, \ldots, r$ are Lagrange multipliers.

The optimization is performed by setting the partial derivative of $L(\tilde{\varphi}_{r+1})$ with respect to $\tilde{\varphi}_{r+1}$ equal to zero

$$2\mathbf{B}\tilde{\varphi}_{r+1} - 2\lambda \mathbf{V} \tilde{\varphi}_{r+1} - \sum_{i=1}^{r} \mu_i \mathbf{R} \varphi_i = \mathbf{0}. \quad \text{(B.4)}$$

Multiplying the left-hand side of (B.4) by $\tilde{\varphi}_{r+1}^T$, and from (B.2), we obtain that

$$\lambda = \frac{\tilde{\varphi}_{r+1}^T \mathbf{B} \tilde{\varphi}_{r+1}}{\tilde{\varphi}_{r+1}^T \mathbf{V} \tilde{\varphi}_{r+1}} = F(\tilde{\varphi}_{r+1}). \quad \text{(B.5)}$$

Thus, λ represents the expression $F(\varphi_{r+1})$ to be maximized.

Multiplying the left-hand side of (B.4) by $\varphi_j^T \mathbf{R}\mathbf{V}^{-1}, j = 1, 2, \ldots, r$, we obtain a set of r expression

$$2\varphi_j^T \mathbf{R}\mathbf{V}^{-1}\mathbf{B}\tilde{\varphi}_{r+1} - \sum_{i=1}^{r} \mu_i \varphi_j^T \mathbf{R}\mathbf{V}^{-1}\mathbf{R}\varphi_i = 0 \quad (j = 1, 2, \ldots, r)$$

that is

$$2\begin{bmatrix}\varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_r^T\end{bmatrix} \mathbf{R}\mathbf{V}^{-1}\mathbf{B}\tilde{\varphi}_{r+1} - \begin{bmatrix}\varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_r^T\end{bmatrix} \mathbf{R}\mathbf{V}^{-1}\mathbf{R} \begin{bmatrix}\varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_r^T\end{bmatrix}^T \begin{bmatrix}\mu_1 \\ \mu_2 \\ \vdots \\ \mu_r\end{bmatrix} = \mathbf{0}. \quad \text{(B.6)}$$

Let

$$\boldsymbol{\mu} = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_r]^T. \quad \text{(B.7)}$$

Using the matrix notation in (10), (B.6) can be represented by

$$\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{R}\mathbf{D}^T\boldsymbol{\mu} = 2\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{B}\tilde{\varphi}_{r+1}. \quad \text{(B.8)}$$

Thus, we obtain

$$\boldsymbol{\mu} = 2(\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{R}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{B}\tilde{\varphi}_{r+1}. \quad \text{(B.9)}$$

It is obvious that the following equation holds:

$$\sum_{i=1}^{r} \mu_i \mathbf{R}\varphi_i = \mathbf{R}\mathbf{D}^T\boldsymbol{\mu}. \quad \text{(B.10)}$$

Thus, (B.4) can be rewritten as the following form:

$$2\mathbf{B}\tilde{\varphi}_{r+1} - 2\lambda\mathbf{V}\tilde{\varphi}_{r+1} - \mathbf{R}\mathbf{D}^T\boldsymbol{\mu} = \mathbf{0}. \quad \text{(B.11)}$$

Substituting (B.9) into (B.11), we have

$$2\mathbf{B}\tilde{\varphi}_{r+1} - 2\lambda\mathbf{V}\tilde{\varphi}_{r+1} - \mathbf{R}\mathbf{D}^T[2(\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{R}\mathbf{D}^T)^{-1} \times \mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{B}\tilde{\varphi}_{r+1}] = \mathbf{0} \quad \text{(B.12)}$$

or equivalently

$$\mathbf{B}\tilde{\varphi}_{r+1} - \mathbf{R}\mathbf{D}^T(\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{R}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{B}\tilde{\varphi}_{r+1} = \lambda\mathbf{V}\tilde{\varphi}_{r+1} \quad \text{(B.13)}$$

that is

$$[\mathbf{I} - \mathbf{R}\mathbf{D}^T(\mathbf{D}\mathbf{R}\mathbf{V}^{-1}\mathbf{R}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{R}\mathbf{V}^{-1}]\mathbf{B}\tilde{\varphi}_{r+1} = \lambda\mathbf{V}\tilde{\varphi}_{r+1}. \quad \text{(B.14)}$$

Considering the matrix notation (9), (B.14) can be expressed as

$$\mathbf{P}\mathbf{B}\tilde{\varphi}_{r+1} = \lambda\mathbf{V}\tilde{\varphi}_{r+1}. \quad \text{(B.15)}$$

Because that $\tilde{\varphi}_{r+1}$ is the direction of φ_{r+1} we, therefore, obtain

$$\mathbf{P}\mathbf{B}\varphi_{r+1} = \lambda\mathbf{V}\varphi_{r+1}. \quad \text{(B.16)}$$

Noting that φ_{r+1} is unit vector, thus, we have

$$\varphi_{r+1}^T\varphi_{r+1} = 1. \quad \text{(B.17)}$$

From (B.16) and (B.17), we can obtain that φ_{r+1} is the eigenvector corresponding to the largest eigenvalue λ of the eigenequation (B.16). □

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and advice.

REFERENCES

[1] J. W. Sammon Jr., "An optimal discriminant plane," *IEEE Trans. Comput.*, vol. C-19, no. 9, pp. 826–829, Sep. 1970.
[2] D. H. Foley and J. W. Sammon Jr., "An optimal set of discriminant vectors," *IEEE Trans. Comput.*, vol. C-24, no. 3, pp. 281–289, Mar. 1975.
[3] T. Okada and S. Tomita, "An optimal orthonormal system for discriminant analysis," *Pattern Recognit.*, vol. 18, no. 2, pp. 139–144, 1985.
[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annu. Eugenics*, vol. 7, pp. 179–188, 1936.
[5] K. Liu, Y. Q. Cheng, J. Y. Yang, and X. Liu, "An efficient algorithm for Foley–Sammon optimal set of discriminant vectors by algebraic method," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 6, no. 5, pp. 817–829, 1992.
[6] Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, pp. 1405–1416, 2001.
[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.
[8] L.-F. Chen, X. Hong-Yuan, M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, pp. 1713–1726, 2000.
[9] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computat.*, vol. 12, pp. 2385–2404, 2000.
[10] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computat.*, vol. 10, pp. 1299–1319, 1998.
[11] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[12] A. Ruiz and P. E. López-de-Teruel, "Nonlinear kernel-based statistical pattern analysis," *IEEE Trans. Neural Netw.*, vol. 12, no. 1, pp. 16–32, Jan. 2001.
[13] K.-R. Müller, S. Mika, G. Rätch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.
[14] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[15] Z. Jin, J.-Y. Yang, Z.-M. Tang, and Z.-S. Hu, "A theorem on the uncorrelated optimal discriminant vectors," *Pattern Recognit.*, vol. 34, pp. 2041–2047, 2001.
[16] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. Piscataway, NJ: IEEE Press, 1999, pp. 41–48.
[17] J. Duchene and A. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 6, pp. 978–983, Jun. 1988.
[18] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 117–126, Jan. 2003.
[19] M.-H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods," in *Proc. 5th Int. Conf. Automatic Face and Gesture Recognition*, Washington, DC, May 2002, pp. 215–220.
[20] J.-T. Chien and C.-C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1644–1649, Dec. 2002.
[21] W. Zheng, L. Zhao, and C. Zou, "An efficient algorithm to solve the small sample size problem for LDA," *Pattern Recognit.*, vol. 37, no. 5, pp. 1077–1079, 2004.
[22] ——, "A modified algorithm for generalized discriminant analysis," *Neural Computat.*, vol. 16, no. 6, pp. 1283–1297, 2004.

Wenming Zheng received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997 and the M.S. degree from Huaqiao University, Quanzhou, China, in 2001. He is currently working toward the Ph.D. degree in the Department of Radio Engineering, Southeast University, Nanjing, China.

His major research interests include neural computation, pattern recognition, machine learning, and computer vision.

Li Zhao was born in Nanjing, China, in 1958. He received the B.E. degree from Nanjing University of Aeronautics and Astronautics, China, in 1982, the M.S. degree from Suzhou University, China, in 1988, and the Ph.D. degree from Kyoto Institute of Technology, Japan, in 1998.

He is currently a Professor with the Department of Radio Engineering, Southeast University, Nanjing, Jiangsu, China. His major interests include pattern recognition and signal processing.

Cairong Zou (M'00) was born in 1963 in Kunshan, Jiangsu, China. He received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from the Department of Radio Engineering, Southeast University, Nanjing, Jiangsu, China, in 1984, 1987, and 1991, respectively.

In 1992, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, Concordia University, Canada. Since 1993, he has been with Southeast University as an Associate Professor. Currently, he is a Professor in the Department of Radio Engineering, Southeast University.