A Rank-One Update Algorithm for Fast Solving Kernel Foley–Sammon Optimal Discriminant Vectors

Wenming Zheng, *Member, IEEE*, Zhouchen Lin, *Senior Member, IEEE*, and Xiaoou Tang, *Fellow, IEEE*

***Abstract*—Discriminant analysis plays an important role in statistical pattern recognition. A popular method is the Foley–Sammon optimal discriminant vectors (FSODVs) method, which aims to find an optimal set of discriminant vectors that maximize the Fisher discriminant criterion under the orthogonal constraint. The FSODVs method outperforms the classic Fisher linear discriminant analysis (FLDA) method in the sense that it can solve more discriminant vectors for recognition. Kernel Foley–Sammon optimal discriminant vectors (KFSODVs) is a nonlinear extension of FSODVs via the kernel trick. However, the current KFSODVs algorithm may suffer from the heavy computation problem since it involves computing the inverse of matrices when solving each discriminant vector, resulting in a cubic complexity for each discriminant vector. This is costly when the number of discriminant vectors to be computed is large. In this paper, we propose a fast algorithm for solving the KFSODVs, which is based on rank-one update (ROU) of the eigensytems. It only requires a square complexity for each discriminant vector. Moreover, we also generalize our method to efficiently solve a family of optimally constrained generalized Rayleigh quotient (OCGRQ) problems which include many existing dimensionality reduction techniques. We conduct extensive experiments on several real data sets to demonstrate the effectiveness of the proposed algorithms.**

***Index Terms*—Dimensionality reduction, discriminant analysis, kernel Foley–Sammon optimal discriminant vectors (KFSODVs), principal eigenvector.**

I. Introduction

DISCRIMINANT analysis is a very active research topic in statistical pattern recognition community. It has been widely used in face recognition [1], [30], image retrieval [2], and text classification [3]. The classical discriminant analysis method is Fisher's linear discriminant analysis (FLDA), which was originally introduced by Fisher [4] for two-class discrimination problems and was further generalized by Rao [5] for multiclass discrimination problems. The basic idea of FLDA is to find an optimal feature space based on Fisher's criterion, namely the projection of the training data onto this space has the maximum ratio of the between-class distance to the within-class distance. For c-class discriminating problems, however, there always exists a so-called *rank limitation* problem, i.e., the rank of the between-class scatter matrix is always bounded by $c-1$, where c is the number of classes. Due to the rank limitation, the maximal number of discriminant vectors of FLDA is $c-1$. However, the $c-1$ discriminant vectors are often insufficient for achieving the best discriminant performance when c is relatively small. This is because some discriminant features fail to be extracted, resulting in a loss of useful discriminant information.

To overcome the rank limitation of FLDA, Foley and Sammon [6] imposed an orthogonal constraint of the discriminant vectors on Fisher's criterion and obtained an optimal set of orthonormal discriminant vectors, to which we refer as the Foley–Sammon optimal discriminant vectors (FSODVs). Okada and Tomita [7] proposed an algorithm based on subspace decomposition to solve FSODVs for multiclass problems, while Duchene and Leclercq [8] proposed an analytic method based on Lagrange multipliers, denoted by FSODVs/LM for simplicity. Although the multiclass FSODVs method overcomes the rank limitation of FLDA, this method can only extract the linear features of the input patterns, and may fail for nonlinear patterns. To this end, in the preliminary work, we have proposed a nonlinear extension of the FSODVs method, called the KFSODVs/LM method for simplicity, by utilizing the kernel trick [9]. We have shown that the KFSODVs method is superior to the FSODVs method in face recognition. However, the current KFSODVs algorithm may suffer from the heavy computation problem since it involves computing the inverse of matrices when solving each discriminant vector, resulting in a cubic complexity for each discriminant vector. This is costly when the number of discriminant vectors to be computed is large.

To reduce the computation cost, in this paper, we propose a fast algorithm for the KFSODVs method. The proposed algorithm is fast and efficient by making use of a rank-one update (ROU) technique, called KFSODVs/ROU algorithm for simplicity, to incrementally establish the eigensystems for the discriminant vectors. To make the ROU technique possible, we elaborate to reformulate the KFSODVs/LM algorithm to an equivalent formulation and further apply the QR decomposition of matrices. Compared with the previous KFSODVs/LM algorithm [9], the new algorithm only requires a square complexity for solving each discriminant vector. To further reduce the complexity of KFSODVs/ROU algorithm, we adopt the modified

Manuscript received June 08, 2008; accepted August 29, 2009. First published January 19, 2010; current version published February 26, 2010. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grants 60503023 and 60872160, and in part by the Science Technology Foundation of Southeast University of China under Grant XJ2008320.

W. Zheng is with the Key Laboratory of Child Development and Learning Science, Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, Jiangsu 210096, China (e-mail: wenming_zheng@seu.edu.cn).

Z. Lin is with Microsoft Research Asia, Beijing 100080, China (e-mail: zhoulin@microsoft.com).

X. Tang was with is with Microsoft Research Asia, Beijing 100080, China. He is now with the Chinese University of Hong Kong, Hong Kong (e-mail: xtang@ie.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2009.2037149

kernel Gram–Schmidt orthogonalization (MKGS) method [10] to replace the kernel principal component analysis (KPCA) [12] method in the preprocessing stage of solving KFSODVs. Moreover, we also extend our algorithm to solve a family of optimally constrained generalized Rayleigh quotient (OCGRQ) problems. Many current dimensionality reduction techniques, such as uncorrelated linear discriminant analysis (ULDA) [13], orthogonal locality preserving projection (OLPP) [14], and those methods in [15]–[19], can be unified into the framework of OCGRQ, hence can be efficiently solved by our algorithm.

The remainder of this paper is organized as follows. In Section II, we briefly review the KFSODVs/LM method. In Section III, we present the fast KFSODVs/ROU algorithm. In Section IV, we propose the fast algorithm for solving the OCGRQ problems. The experiments are presented in Section V. Finally, Section VI concludes our paper.

II. Brief Review of KFSODVs

Let $\{\mathbf{x}_i | i = 1, 2, \ldots, n\}$ be a set of n d-dimensional samples and l_i be the class label of $\mathbf{x}_i$, where $l_i \in \{1, 2, \ldots, c\}$ and c is the number of classes. Denote the number of the ith class samples by n_i. Let Φ be a nonlinear mapping which maps the input space into a high-dimensional feature space $\mathcal{F}$, that is

$$\Phi : \mathbf{x}_i \rightarrow \Phi(\mathbf{x}_i).$$

Then, the between-class, the within-class, and the total-class scatter matrices in the feature space are defined as

$$\begin{aligned}
\mathbf{S}_b^\Phi &= \sum_{i=1}^{c} n_i (\mathbf{m}_i^\Phi - \mathbf{m}^\Phi)(\mathbf{m}_i^\Phi - \mathbf{m}^\Phi)^T = \mathbf{H}_b^\Phi {\mathbf{H}_b^\Phi}^T \\
\mathbf{S}_w^\Phi &= \sum_{i=1}^{n} (\Phi(\mathbf{x}_i) - \mathbf{m}_{l_i}^\Phi)(\Phi(\mathbf{x}_i) - \mathbf{m}_{l_i}^\Phi)^T = \mathbf{H}_w^\Phi {\mathbf{H}_w^\Phi}^T \\
\mathbf{S}_t^\Phi &= \sum_{i=1}^{n} (\Phi(\mathbf{x}_i) - \mathbf{m}^\Phi)(\Phi(\mathbf{x}_i) - \mathbf{m}^\Phi)^T = \mathbf{H}_t^\Phi {\mathbf{H}_t^\Phi}^T \quad (1)
\end{aligned}$$

respectively, where $\mathbf{m}_i^\Phi = 1/n_i \sum_{j:l_j=i} \Phi(\mathbf{x}_j)$ is the mean of the ith class, $\mathbf{m}^\Phi = 1/n \sum_{i=1}^{n} \Phi(\mathbf{x}_i)$ is the mean of all samples, and $\mathbf{H}_b^\Phi$, $\mathbf{H}_w^\Phi$, and $\mathbf{H}_t^\Phi$ are, respectively, defined as shown in the equation at the bottom of the page. Then, the Fisher's discriminant criterion in the feature space $\mathcal{F}$ can be expressed as

$$J^\Phi(\omega) = \frac{\omega^T \mathbf{S}_b^\Phi \omega}{\omega^T \mathbf{S}_w^\Phi \omega}. \quad (2)$$

The basic idea of the KFSODVs method is to find an optimal set of discriminant vectors, denoted by $\omega_1^\Phi, \omega_2^\Phi, \ldots, \omega_r^\Phi$, that maximize $J^\Phi(\omega)$ under the orthogonal constraints

$$\left(\omega_i^\Phi\right)^T \omega_j^\Phi = 0, \qquad i \neq j. \quad (3)$$

More specifically, the optimal set of discriminant vectors of the KFSODVs method can be successively computed by solving the following sequence of optimization problems:

$$\begin{aligned}
\omega_1^\Phi &= \arg\max_{\omega} \frac{\omega^T \mathbf{S}_b^\Phi \omega}{\omega^T \mathbf{S}_w^\Phi \omega} \\
&\cdots \\
\omega_r^\Phi &= \arg \max_{\substack{\omega^T \omega_i = 0, \\ i=1,2,\ldots,r-1}} \frac{\omega^T \mathbf{S}_b^\Phi \omega}{\omega^T \mathbf{S}_w^\Phi \omega}. \quad (4)
\end{aligned}$$

In [9], we have proposed the KFSODVs/LM algorithm to solve the KFSODVs by using the Lagrangian multipliers, where we divide the whole feature space into three subspaces, i.e., $\mathbf{S}_t^\Phi(0)$, $\overline{\mathbf{S}_t^\Phi(0)} \cap \mathbf{S}_w^\Phi(0)$, and $\overline{\mathbf{S}_t^\Phi(0)} \cap \overline{\mathbf{S}_w^\Phi(0)}$, in which $\mathbf{A}(0)$ is the null space of $\mathbf{A}$ and $\overline{\mathbf{B}}$ is the orthogonal complement of $\mathbf{B}$. Because the subspace $\mathbf{S}_t^\Phi(0)$ contains no useful discriminant information [9], we solve the discriminant vectors in the later two subspaces. The solution procedures of the KFSODVs/LM algorithm can be expressed as follows.

1) Solve the orthonormal basis of the subspace $\overline{\mathbf{S}_t^\Phi(0)}$ via KPCA and denote $\mathbf{W}_{\mathrm{KPCA}}^\Phi$ as the KPCA transform matrix.
2) Project the input data $\Phi(\mathbf{x}_i)$ onto $\mathbf{W}_{\mathrm{KPCA}}^\Phi$, that is, $\mathbf{y}_i = \left(\mathbf{W}_{\mathrm{KPCA}}^\Phi\right)^T \Phi(\mathbf{x}_i)$.
3) In the data set $\{\mathbf{y}_i | i = 1, 2, \ldots, n\}$, compute the within-class scatter matrix

$$\mathbf{S}_w^y = \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{m}_{l_i}^y)(\mathbf{y}_i - \mathbf{m}_{l_i}^y)^T = \mathbf{H}_w^y {\mathbf{H}_w^y}^T$$

where $\mathbf{m}_i^y = 1/n_i \sum_{j:l_j=i} \mathbf{y}_j$ and

$$\mathbf{H}_w^y = (\mathbf{y}_1 - \mathbf{m}_{l_1}^y \quad \mathbf{y}_2 - \mathbf{m}_{l_2}^y \quad \cdots \quad \mathbf{y}_n - \mathbf{m}_{l_n}^y).$$

4) Perform the singular value decomposition (SVD) on $\mathbf{S}_w^y$ to find the orthonormal basis of $\mathbf{S}_w^y(0)$ and $\overline{\mathbf{S}_w^y(0)}$, respectively. Denote $\mathbf{W}_y$ and $\mathbf{W}_y^\perp$ as the transformation matrices whose columns consist of the orthonormal basis of $\mathbf{S}_w^y(0)$ and $\overline{\mathbf{S}_w^y(0)}$, respectively, and compute $\mathbf{z}_i = \mathbf{W}_y^T \mathbf{y}_i$ and $\mathbf{z}_i^\perp = \left(\mathbf{W}_y^\perp\right)^T \mathbf{y}_i$.
5) In the data set $\{\mathbf{z}_i | i = 1, 2, \ldots, n\}$, compute the between-class scatter matrix

$$\mathbf{S}_b^z = \sum_{i=1}^{c} n_i (\mathbf{m}_i^z - \mathbf{m}^z)(\mathbf{m}_i^z - \mathbf{m}^z)^T = \mathbf{H}_b^z \left(\mathbf{H}_b^z\right)^T$$

where $\mathbf{m}_i^z = 1/n_i \sum_{j:l_j=i} \mathbf{z}_j$, $\mathbf{m}^z = 1/n \sum_{i=1}^{n} \mathbf{z}_i$, and $\mathbf{H}_b^z = (\sqrt{n_1}(\mathbf{m}_1^z - \mathbf{m}^z) \quad \sqrt{n_2}(\mathbf{m}_2^z - \mathbf{m}^z) \quad \cdots \quad \sqrt{n_c}(\mathbf{m}_c^z - \mathbf{m}^z))$.

$$\begin{aligned}
\mathbf{H}_b^\Phi &= (\sqrt{n_1}(\mathbf{m}_1^\Phi - \mathbf{m}^\Phi) \quad \sqrt{n_2}(\mathbf{m}_2^\Phi - \mathbf{m}^\Phi) \quad \cdots \quad \sqrt{n_c}(\mathbf{m}_c^\Phi - \mathbf{m}^\Phi)) \\
\mathbf{H}_w^\Phi &= (\Phi(\mathbf{x}_1) - \mathbf{m}_{l_1}^\Phi \quad \Phi(\mathbf{x}_2) - \mathbf{m}_{l_2}^\Phi \quad \cdots \quad \Phi(\mathbf{x}_n) - \mathbf{m}_{l_n}^\Phi) \\
\mathbf{H}_t^\Phi &= (\Phi(\mathbf{x}_1) - \mathbf{m}^\Phi \quad \Phi(\mathbf{x}_2) - \mathbf{m}^\Phi \quad \cdots \quad \Phi(\mathbf{x}_n) - \mathbf{m}^\Phi).
\end{aligned}$$

6) Similarly, in the data set $\{\mathbf{z}_i^{\perp} | i = 1, 2, \ldots, n\}$, compute the between-class and within-class scatter matrices

$$\mathbf{S}_b^{z^\perp} = \sum_{i=1}^{c} n_i (\mathbf{m}_i^{z^\perp} - \mathbf{m}^{z^\perp})(\mathbf{m}_i^{z^\perp} - \mathbf{m}^{z^\perp})^T = \mathbf{H}_b^{z^\perp} \left(\mathbf{H}_b^{z^\perp}\right)^T$$

and

$$\mathbf{S}_w^{z^\perp} = \sum_{i=1}^{n} (\mathbf{z}_i^{\perp} - \mathbf{m}_{l_i}^{z^\perp})(\mathbf{z}_i^{\perp} - \mathbf{m}_{l_i}^{z^\perp})^T = \mathbf{H}_w^{z^\perp} \left(\mathbf{H}_w^{z^\perp}\right)^T$$

where $\mathbf{m}_i^{z^\perp} = 1/n_i \sum_{j:l_j=i} \mathbf{z}_j^{\perp}$, $\mathbf{m}^{z^\perp} = 1/n \sum_{i=1}^{n} \mathbf{z}_i^{\perp}$, $\mathbf{H}_b^{z^\perp} = (\sqrt{n_1}(\mathbf{m}_1^{z^\perp} - \mathbf{m}^{z^\perp}) \quad \sqrt{n_2}(\mathbf{m}_2^{z^\perp} - \mathbf{m}^{z^\perp}) \quad \cdots \quad \sqrt{n_c}(\mathbf{m}_c^{z^\perp} - \mathbf{m}^{z^\perp}))$, and $\mathbf{H}_w^{z^\perp} = (\mathbf{z}_1^{\perp} - \mathbf{m}_{l_1}^{z^\perp} \quad \mathbf{z}_2^{\perp} - \mathbf{m}_{l_2}^{z^\perp} \quad \cdots \quad \mathbf{z}_n^{\perp} - \mathbf{m}_{l_n}^{z^\perp})$.

7) Solve the principal eigenvectors $\phi_1, \phi_2, \ldots, \phi_l$ of the following eigensystem:

$$\mathbf{S}_b^{z}\phi = \lambda\phi. \tag{5}$$

Then

$$\omega_i^{\Phi} = \mathbf{W}_{\mathrm{KPCA}}^{\Phi} \mathbf{W}_y \phi_i, \qquad i = 1, 2, \ldots, l$$

are the optimal discriminant vectors of KFSODVs lying in the subspace $\mathbf{S}_w^{\Phi}(0)$.

8) Assume that $\omega_1, \omega_2, \ldots, \omega_k$ are the first k optimal discriminant vectors maximizing the Fisher's discriminant criterion $J(\omega) = \omega^T \mathbf{S}_b^{z^\perp} \omega / \omega^T \mathbf{S}_w^{z^\perp} \omega$ under the orthogonal constraints $\omega_i^T \omega_j = 0$ $(i \neq j)$. Then

$$\omega_{l+i}^{\Phi} = \mathbf{W}_{\mathrm{KPCA}}^{\Phi} \mathbf{W}_y^{\perp} \omega_i, \qquad i = 1, 2, \ldots, k$$

are the first k optimal discriminant vectors of KFSODVs lying in the subspace $\overline{\mathbf{S}_w^{\Phi}(0)}$, where the k discriminant vectors $\omega_1, \omega_2, \ldots, \omega_k$ can be solved by the following procedures.

- The first optimal discriminant vector ω_1 is the eigenvector associated with the largest eigenvalue of the eigensystem

$$(\mathbf{S}_w^{z^\perp})^{-1} \mathbf{S}_b^{z^\perp} \omega = \lambda\omega. \tag{6}$$

- Suppose that the first r optimal discriminant vectors $\omega_1, \omega_2, \ldots, \omega_r$ have been obtained, then the $(r+1)$th optimal discriminant vector ω_{r+1} is the eigenvector associated with the largest eigenvalue of the eigensystem

$$(\mathbf{I}_m - (\mathbf{S}_w^{z^\perp})^{-1} \mathbf{D}_r (\mathbf{D}_r^T (\mathbf{S}_w^{z^\perp})^{-1} \mathbf{D}_r)^{-1} \mathbf{D}_r^T) \times (\mathbf{S}_w^{z^\perp})^{-1} \mathbf{S}_b^{z^\perp} \omega = \lambda\omega \tag{7}$$

where $\mathbf{D}_r = (\omega_1 \quad \omega_2 \quad \cdots \quad \omega_r)$ and $\mathbf{I}_m$ is the $m \times m$ identity matrix, where m represents the dimensionality of $\mathbf{z}_i^{\perp}$ $(i = 1, 2, \ldots, n)$.

From the above procedures, one can see that the most time consuming part of solving the KFSODVs is to solve the k optimal discriminant vectors in subspace $\overline{\mathbf{S}_w^{\Phi}(0)}$, especially when k is large. The most preferred method for computing the principal eigenvector of both (6) and (7) is the power method [20], where only a square complexity is needed to solve the principal eigenvector of the eigensystem. However, for (7), after computing the matrix $\mathbf{D}_r^T (\mathbf{S}_w^{z^\perp})^{-1} \mathbf{D}_r$, one also has to compute its inverse when computing each new discriminant vector. Moreover, it should be noted that using the KPCA method to find the orthogonal basis of $\overline{\mathbf{S}_t^{\Phi}(0)}$ in step 1) may also be time consuming. To this end, we also adopt the MKGS method to replace the KPCA method for finding the basis of the principal subspace of the total-class scatter matrix, which further reduces the computation cost of our algorithm.

Algorithm 1 lists the pseudocode of solving the k optimal discriminant vectors defined in (6) and (7). Table I shows the computation complexity of Algorithm 1. From Table I, one can see that there will be a complexity of $O(r^3 + m^2 r)$ for the $(r+1)$th discriminant vector, where m is the dimensionality of $\overline{\mathbf{S}_w^{y}(0)}$. Therefore, the total computational complexity of Algorithm 1 could be as large as $O(m^2 n + m^3 + k^4 + m^2 k^2)$ if we want to obtain k discriminant vectors. Hence, the computation of KFSODVs is particularly costly when k becomes large.

Algorithm 1: KFSODVs/LM algorithm for solving the k optimal vectors $\omega_1, \ldots, \omega_k$ in (6) and (7)

Input:

- Data matrices $\mathbf{H}_b^{z^\perp} = (\sqrt{n_1}(\mathbf{m}_1^{z^\perp} - \mathbf{m}^{z^\perp}) \quad \sqrt{n_2}(\mathbf{m}_2^{z^\perp} - \mathbf{m}^{z^\perp}) \quad \cdots \quad \sqrt{n_c}(\mathbf{m}_c^{z^\perp} - \mathbf{m}^{z^\perp}))$ and $\mathbf{H}_w^{z^\perp} = (\mathbf{z}_1^{\perp} - \mathbf{m}_{l_1}^{z^\perp} \quad \mathbf{z}_2^{\perp} - \mathbf{m}_{l_2}^{z^\perp} \quad \cdots \quad \mathbf{z}_n^{\perp} - \mathbf{m}_{l_n}^{z^\perp})$, label vector $\mathbf{l} = (l_1 \quad l_2 \quad \cdots \quad l_n)^T$, and the number k of optimal discriminant vectors.

Initialization:

1) Compute $\mathbf{S}_w^{z^\perp} = \mathbf{H}_w^{z^\perp} (\mathbf{H}_w^{z^\perp})^T$.
2) Compute the inverse of $\mathbf{S}_w^{z^\perp}$, i.e., $(\mathbf{S}_w^{z^\perp})^{-1}$.
3) Compute the principal eigenvector of $(\mathbf{S}_w^{z^\perp})^{-1} \mathbf{H}_b^{z^\perp} (\mathbf{H}_b^{z^\perp})^T \omega = \lambda\omega$ via the power method and set $\mathbf{D}_1 \leftarrow \omega$.

For $r = 2, \ldots, k$, **Do**

1) Compute $(\mathbf{D}_{r-1}^T (\mathbf{S}_w^{z^\perp})^{-1} \mathbf{D}_{r-1})^{-1}$.
2) Compute the principal eigenvector of

$$(\mathbf{I}_m - (\mathbf{S}_w^{z^\perp})^{-1} \mathbf{D}_{r-1} (\mathbf{D}_{r-1}^T (\mathbf{S}_w^{z^\perp})^{-1} \mathbf{D}_{r-1})^{-1} \mathbf{D}_{r-1}^T) \times (\mathbf{S}_w^{z^\perp})^{-1} \mathbf{H}_b^{z^\perp} \left(\mathbf{H}_b^{z^\perp}\right)^T \omega = \lambda\omega$$

via the power method.

3) Set $\mathbf{D}_r \leftarrow (\mathbf{D}_{r-1} \quad \omega)$.

Output:

- Output the columns of $\mathbf{D}_k = [\omega_1, \omega_2, \ldots, \omega_k]$.

TABLE I
COMPUTATIONAL COMPLEXITY OF ALGORITHM 1

Algorithm 1	Initialization Part			The r-th Loop	
Line No.	1)	2)	3)	1)	2)
Computational Complexity	$O(m^2n)$	$O(m^3)$	$O(m^2)$	$O(m^2r+r^3)$	$O(m^2)$

III. KFSODVS/ROU ALGORITHM FOR FAST SOLVING KFSODVS

We have reviewed the KFSODVs/LM algorithm in Section II, from which one can see that the KFSODVs/LM algorithm can be expressed in the following form:

$$\text{KFSODVs/LM} = \text{KPCA} + \text{FSODVs/LM}$$

where KPCA is used to find the orthonormal basis of the subspace $\overline{\mathbf{S}_t^\Phi(0)}$.

In this section, we will present the KFSODVs/ROU algorithm to efficiently solve the discriminant vectors of the KFSODVs method. To further reduce the computational cost, we adopt the MKGS algorithm to find the orthonormal basis of $\overline{\mathbf{S}_t^\Phi(0)}$ rather than using the KPCA algorithm. Moreover, to efficiently solve the optimal set of discriminant vectors in (6) and (7), we propose to make use of the ROU technique to incrementally establish the eigensystems for solving these discriminant vectors. Consequently, our KFSODVs/ROU algorithm can be expressed in the following form:

$$\text{KFSODVs/ROU} = \text{MKGS} + \text{FSODVs/ROU}. \tag{8}$$

A. MKGS for Solving the Orthonormal Basis of $\overline{\mathbf{S}_t^\phi(0)}$

Let $\mathbf{h}_i^\Phi = \Phi(\mathbf{x}_i) - \mathbf{m}^\Phi$. Then, the matrix $\mathbf{H}_t^\Phi$ defined in (1) can be expressed as $\mathbf{H}_t^\Phi = \left[\mathbf{h}_1^\Phi, \ldots, \mathbf{h}_n^\Phi\right]$. According to the MKGS algorithm [10], the corresponding orthonormal vectors of the columns of $\mathbf{H}_t^\Phi$ can be obtained using the following steps, where $\mathbf{s}_j$ and $\mathbf{t}_j$ $(j = 1, \ldots, n)$ are n-dimensional vectors, $\mathbf{D}$ is an $n \times n$ diagonal matrix, $\mathbf{\Delta}$ is an $n \times n$ matrix, $\mathbf{e}_j = (0, \ldots, 1, \ldots, 0)^T$ is an n-dimensional vector where the jth item is 1, and $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the inner product of vectors $\mathbf{x}$ and $\mathbf{y}$.

1) Let $\mathbf{s}_1 = \mathbf{t}_1 = \mathbf{e}_1$, $\mathbf{D}_{11} = \langle \mathbf{h}_1^\Phi, \mathbf{h}_1^\Phi \rangle$,[1] and $\mathbf{\Delta}_{i1} = \langle \mathbf{h}_i^\Phi, \mathbf{h}_1^\Phi \rangle$ $(i = 1, \ldots, n)$.
2) Repeat for $j = 2, \ldots, n$:
   a) $\mathbf{t}_j^{(1)} = \mathbf{e}_j$;
   b) repeat for $i = 1, \ldots, j-1$:
      i) $\mathbf{s}_{ji} = \sum_{p=1}^{j} \mathbf{\Delta}_{pi} \mathbf{t}_{pj}^{(i)} / \mathbf{D}_{ii}$;
      ii) $\mathbf{t}_j^{i+1} = \mathbf{t}_j^{(i)} - s_{ji}\mathbf{t}_i$;
   c) $\mathbf{t}_j = \mathbf{t}_j^{(j)}$;
   d) repeat for $p = 1, \ldots, n$:
      i) $\mathbf{\Delta}_{pj} = \sum_{q=1}^{j} \mathbf{t}_{qj} \langle \mathbf{h}_q^\Phi, \mathbf{h}_p^\Phi \rangle$;
   e) compute $\mathbf{D}_{jj} = \sum_{p=1}^{j} \mathbf{\Delta}_{pj} \mathbf{t}_{pj}$.
3) $\mathbf{R} = \mathbf{D}^{1/2}[\mathbf{s}_1, \ldots, \mathbf{s}_n]$, where $\mathbf{s}_i = [s_{i1}, s_{i2}, \ldots, s_{i(i-1)}, 1, 0, \ldots, 0]^T$.
4) $\mathbf{R}^{-1} = [\mathbf{t}_1, \ldots, \mathbf{t}_n]\,\mathbf{D}^{-1/2}$.

[1]The inner product $\langle \mathbf{h}_i^\Phi, \mathbf{h}_j^\Phi \rangle$ can be computed via the kernel trick.

Then the columns of the matrix $\mathbf{H}_t^\Phi[\mathbf{t}_1, \ldots, \mathbf{t}_n]\,\mathbf{D}^{-1/2}$ are the corresponding orthonormal vectors of the columns of $\mathbf{H}_t^\Phi$.

If the column space of $\mathbf{H}_t^\Phi$ is not of full rank, i.e., $\text{rank}(\mathbf{H}_t^\Phi) < n$, then there are $n - \text{rank}(\mathbf{H}_t^\Phi)$ diagonal elements of $\mathbf{D}$ that would vanish. In this case, we should omit those $\mathbf{t}_j$ for which $\mathbf{D}_{jj} = 0$ and obtain $\mathbf{R}^{-1}$ whose number of columns is equal to the rank of $\mathbf{H}_t^\Phi$.

B. KFSODVs/ROU Algorithm for Solving the Discriminant Vectors in (6) and (7)

Rewrite the k optimal discriminant vectors in (6) and (7) into the following successive form:

$$\begin{aligned} \omega_1 &= \arg\max_{\omega} \frac{\omega^T \mathbf{H}_b^{z^\perp}\left(\mathbf{H}_b^{z^\perp}\right)^T \omega}{\omega^T \mathbf{S}_w^{z^\perp} \omega} \\ &\cdots \\ \omega_k &= \arg \max_{\substack{\omega^T\omega_i = 0,\\ i=1,2,\ldots,k-1}} \frac{\omega^T \mathbf{H}_b^{z^\perp}\left(\mathbf{H}_b^{z^\perp}\right)^T \omega}{\omega^T \mathbf{S}_w^{z^\perp} \omega}. \end{aligned} \tag{9}$$

Solving (9) boils down to solving the following optimization problem:

$$\begin{aligned} \omega_1 &= \arg \max_{\omega^T \mathbf{S}_w^{z^\perp}\omega = 1} \omega^T \mathbf{H}_b^{z^\perp}\left(\mathbf{H}_b^{z^\perp}\right)^T \omega \\ &\cdots \\ \omega_k &= \arg \max_{\substack{\omega^T \mathbf{S}_w^{z^\perp}\omega = 1,\\ \omega^T\omega_i = 0,\\ i=1,2,\ldots,k-1}} \omega^T \mathbf{H}_b^{z^\perp}\left(\mathbf{H}_b^{z^\perp}\right)^T \omega. \end{aligned} \tag{10}$$

Let $\mathbf{S}_w^{z^\perp} = \mathbf{U}_w \mathbf{\Lambda}_w \mathbf{U}_w^T$ be the SVD of $\mathbf{S}_w^{z^\perp}$, where $\mathbf{U}_w$ is an $m \times m$ orthogonal matrix and $\mathbf{\Lambda}_w = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$. Since $\mathbf{S}_w^{z^\perp}$ is nonsingular, we have $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$. Let

$$\mathbf{L}_w = \mathbf{U}_w \mathbf{\Lambda}_w^{-1/2} \quad \text{and} \quad \omega = \mathbf{L}_w \alpha.$$

Then, solving the optimal discriminant vectors $\omega_1, \omega_2, \ldots, \omega_k$ boils down to solving the optimal vectors $\alpha_1, \alpha_2, \ldots, \alpha_k$ of the following optimization problems:

$$\begin{aligned} \alpha_1 &= \arg \max_{\alpha^T\alpha = 1} \alpha^T \mathbf{L}_w^T \mathbf{H}_b^{z^\perp}\left(\mathbf{H}_b^{z^\perp}\right)^T \mathbf{L}_w \alpha \\ &\cdots \\ \alpha_k &= \arg \max_{\substack{\alpha^T\alpha = 1,\\ \alpha^T \mathbf{\Lambda}_w^{-1}\mathbf{U}_{k-1} = \mathbf{0}}} \alpha^T \mathbf{L}_w^T \mathbf{H}_b^{z^\perp}\left(\mathbf{H}_b^{z^\perp}\right)^T \mathbf{L}_w \alpha \end{aligned} \tag{11}$$

where $\mathbf{U}_{k-1} = (\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_{k-1})$. The first vector α_1 is the principal eigenvector associated with the largest eigenvalue of the matrix $\mathbf{L}_w^T \mathbf{H}_b^{z^\perp}(\mathbf{H}_b^{z^\perp})^T \mathbf{L}_w$. Suppose that we have obtained the first r vectors $\alpha_1, \ldots, \alpha_r$. To solve the $(r+1)$th vector α_{r+1}, we introduce Lemma 1 and Theorems 1 and 2. Similar proofs had been given in [11]. We provide the proofs in Appendixes I–III, respectively.

TABLE II
COMPUTATIONAL COMPLEXITY OF ALGORITHM 2

Algorithm 2	Initialization Part		The r-th Loop			
Line No.	1)	2)	1)	2)	3)	4)
Computational Complexity	$O(m^2n)$	$O(m^2c)$	$O(mc)$	$O(m^2)$	$O(mr)$	$O(mc)$

Lemma 1: Let $\mathbf{Q}$ be an $m \times r$ $(r < m)$ matrix with orthonormal columns. If $\alpha^T\mathbf{Q} = \mathbf{0}$, then there exists a (nonunique) $\beta \in \mathcal{R}^m$ such that

$$\alpha = (\mathbf{I}_m - \mathbf{Q}\mathbf{Q}^T)\beta.$$

Theorem 1: Let $\mathbf{Q}_r\mathbf{R}_r$ be the QR decomposition[2] of $\mathbf{\Lambda}_w^{-1}\mathbf{U}_r$. Then, α_{r+1} is the principal eigenvector corresponding to the largest eigenvalue of the following matrix:

$$(\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{L}_w^T\mathbf{H}_b^{z^\perp}\left(\mathbf{H}_b^{z^\perp}\right)^T\mathbf{L}_w(\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T).$$

Theorem 2: Suppose that $\mathbf{Q}_r\mathbf{R}_r$ is the QR decomposition of $\mathbf{\Lambda}_w^{-1}\mathbf{U}_r$. Let $\mathbf{U}_{r+1} = (\mathbf{U}_r \quad \alpha_{r+1})$, $\mathbf{q} = \mathbf{\Lambda}_w^{-1}\alpha_{r+1} - \mathbf{Q}_r(\mathbf{Q}_r^T\mathbf{\Lambda}_w^{-1}\alpha_{r+1})$, and $\mathbf{Q}_{r+1} = (\mathbf{Q}_r \quad \mathbf{q}/\|\mathbf{q}\|)$. Then

$$\mathbf{Q}_{r+1}\begin{pmatrix}\mathbf{R}_r & \mathbf{Q}_r^T\mathbf{\Lambda}_w^{-1}\alpha_{r+1}\\ \mathbf{0}^T & \|\mathbf{q}\|\end{pmatrix}$$

is the QR decomposition of $\mathbf{\Lambda}_w^{-1}\mathbf{U}_{r+1}$.

The above two theorems are crucial to design our fast algorithm. Theorem 1 makes it possible to use the power method to solve the optimal discriminant vectors, while Theorem 2 makes it possible to update $\mathbf{Q}_{r+1}$ from $\mathbf{Q}_r$ by adding a single column. Moreover, it is notable that

$$\begin{aligned}\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T &= \prod_{i=1}^{r}(\mathbf{I}_m - \mathbf{q}_i\mathbf{q}_i^T)\\ &= (\mathbf{I}_m - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)(\mathbf{I}_m - \mathbf{q}_r\mathbf{q}_r^T) \qquad (12)\end{aligned}$$

where $\mathbf{q}_i$ is the ith column of $\mathbf{Q}_r$. Equation (12) makes it possible to update the positive-semidefinite matrix $(\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{L}_w^T\mathbf{H}_b^{z^\perp}(\mathbf{H}_b^{z^\perp})^T\mathbf{L}_w(\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T)$ from $(\mathbf{I}_m - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)\mathbf{L}_w^T\mathbf{H}_b^{z^\perp}(\mathbf{H}_b^{z^\perp})^T\mathbf{L}_w(\mathbf{I}_m - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)$ by the ROU technique.

Based on the above derivations, we summarize the fast algorithm of solving the optimal discriminant vectors $\omega_1, \omega_2, \ldots, \omega_k$ in Algorithm 2. Note that we compute the SVD of $\mathbf{H}_w^{z^\perp}$ instead of $\mathbf{S}_w^{z^\perp}$ therein in order to save the computation of preparing $\mathbf{S}_w^{z^\perp}$ by computing $\mathbf{H}_w^{z^\perp}(\mathbf{H}_w^{z^\perp})^T$. This $O(m^2n)$ computation cannot be waived in Algorithm 1.

Algorithm 2: KFSODVs/ROU algorithm for solving the k optimal vectors $\omega_1, \ldots, \omega_k$ in (6) and (7)

Input:
- Data matrices $\mathbf{H}_b^{z^\perp} = (\sqrt{n_1}(\mathbf{m}_1^{z^\perp} - \mathbf{m}^{z^\perp}) \quad \sqrt{n_2}(\mathbf{m}_2^{z^\perp} - \mathbf{m}^{z^\perp}) \quad \cdots \quad \sqrt{n_c}(\mathbf{m}_c^{z^\perp} - \mathbf{m}^{z^\perp}))$ and $\mathbf{H}_w^{z^\perp} = (\mathbf{z}_1^\perp - \mathbf{m}_{l_1}^{z^\perp} \quad \mathbf{z}_2^\perp - \mathbf{m}_{l_2}^{z^\perp} \quad \cdots \quad \mathbf{z}_n^\perp - \mathbf{m}_{l_n}^{z^\perp})$, label vector $\mathbf{l} = (l_1 \quad l_2 \quad \cdots \quad l_n)^T$, the number k of optimal discriminant vectors, and the threshold $\varepsilon > 0$.

Initialization:
1) Perform SVD of $\mathbf{H}_w^{z^\perp}$: $\mathbf{H}_w^{z^\perp} = \mathbf{U}_w\mathbf{\Lambda}_w^{1/2}\mathbf{V}_w^T$.
2) $\mathbf{L}_w \leftarrow \mathbf{U}_w\mathbf{\Lambda}_w^{-1/2}$ and $\mathbf{H}_1 \leftarrow \left(\mathbf{H}_b^{z^\perp}\right)^T\mathbf{L}_w$.

For $r = 1, 2, \ldots, k$, **Do**
1) Solve the principal eigenvector of $\mathbf{H}_r^T\mathbf{H}_r\alpha = \lambda\alpha$ via the following power method:
   While $\|\alpha^{(j)} - \alpha^{(j-1)}\| > \varepsilon$, do
   - $\mathbf{y}^{(j)} = \mathbf{H}_r\alpha^{(j-1)}$, $\mathbf{z}^{(j)} = \mathbf{H}_r^T\mathbf{y}^{(j)}$, $\alpha^{(j)} = \mathbf{z}^{(j)}/\|\mathbf{z}^{(j)}\|$;
   - $j = j + 1$.
2) $\omega_r \leftarrow \mathbf{L}_w\alpha$ and $\omega_r \leftarrow \omega_r/\|\omega_r\|$.
3) If $r = 1$, $\mathbf{q}_r \leftarrow \mathbf{\Lambda}_w^{-1}\alpha$, $\mathbf{q}_r \leftarrow \mathbf{q}_r/\|\mathbf{q}_r\|$, and $\mathbf{Q}_1 \leftarrow \mathbf{q}_r$; else $\mathbf{q}_r \leftarrow \mathbf{\Lambda}_w^{-1}\alpha - \mathbf{Q}_{r-1}(\mathbf{Q}_{r-1}^T(\mathbf{\Lambda}_w^{-1}\alpha))$, $\mathbf{q}_r \leftarrow \mathbf{q}_r/\|\mathbf{q}_r\|$, and $\mathbf{Q}_r \leftarrow (\mathbf{Q}_{r-1} \quad \mathbf{q}_r)$.
4) $\mathbf{H}_{r+1} \leftarrow \mathbf{H}_r - (\mathbf{H}_r\mathbf{q}_r)\mathbf{q}_r^T$.

Output:
- Output $\omega_1, \omega_2, \ldots, \omega_k$.

C. Computational Complexity

Table II shows the detailed computational complexity analysis of Algorithm 2. In the initialization part, each operation is performed only once. In the loop part, however, each operation needs to repeat k times if we want to obtain k optimal discriminant vectors. From Table II, one can see that there only needs to be a complexity of $O(m^2)$ for the $(r+1)$th discriminant vector, and the total complexity of solving the k optimal discriminant vectors $\omega_1, \omega_2, \ldots, \omega_k$ is $O\{m^2(n + c + k)\}$. In contrast with Algorithm 1, one can see that the computation cost of KFSODVs/ROU algorithm is significantly reduced.

D. KFSODVs for Recognition

In this section, we will address the recognition problem based on the KFSODVs method. Let $\mathbf{W}_N = \left[\omega_1^\Phi, \ldots, \omega_l^\Phi\right]$ denote the optimal transform matrix obtained from the subspace $\overline{\mathbf{S}_t^\Phi(0)} \cap \mathbf{S}_w^\Phi(0)$ and $\mathbf{W}_C = \left[\omega_{l+1}^\Phi, \ldots, \omega_{l+k}^\Phi\right]$ the optimal transform matrix obtained from the subspace $\overline{\mathbf{S}_t^\Phi(0)} \cap \overline{\mathbf{S}_w^\Phi(0)}$. Suppose that $\Phi(\mathbf{x}_{\text{test}})$ is a test sample. Then, the projections of $\Phi(\mathbf{x}_{\text{test}})$ onto $\mathbf{W}_N$ and $\mathbf{W}_C$ are

$$\mathbf{y}_{\text{test}} = \mathbf{W}_N^T\Phi(\mathbf{x}_{\text{test}}) \quad \text{and} \quad \mathbf{z}_{\text{test}} = \mathbf{W}_C^T\Phi(\mathbf{x}_{\text{test}}). \qquad (13)$$

Let

$$\mathbf{y}_i = \mathbf{W}_N^T\Phi(\mathbf{x}_i) \quad \text{and} \quad \mathbf{z}_i = \mathbf{W}_C^T\Phi(\mathbf{x}_i), \qquad i = 1, \ldots, n \qquad (14)$$

be the projections of $\Phi(\mathbf{x}_i)$ onto $\mathbf{W}_N$ and $\mathbf{W}_C$, respectively. Denote the distance between $\mathbf{y}_{\text{test}}$ and $\mathbf{y}_i$ by $d_N(\mathbf{y}_{\text{test}}, \mathbf{y}_i)$ and the distance between $\mathbf{z}_{\text{test}}$ and $\mathbf{z}_i$ by $d_C(\mathbf{z}_{\text{test}}, \mathbf{z}_i)$, where

$$d_N(\mathbf{y}_{\text{test}}, \mathbf{y}_i) = \|\mathbf{y}_{\text{test}} - \mathbf{y}_i\| \qquad (15)$$

$$d_C(\mathbf{z}_{\text{test}}, \mathbf{z}_i) = \|\mathbf{z}_{\text{test}} - \mathbf{z}_i\|. \qquad (16)$$

Considering that the distances $d_N(\mathbf{y}_{\text{test}}, \mathbf{y}_i)$ and $d_C(\mathbf{z}_{\text{test}}, \mathbf{z}_i)$ are computed from different subspaces, they may not share the

[2]Given an $m \times n$ matrix $\mathbf{A}$, the QR decomposition of $\mathbf{A}$ is to find an $m \times n$ orthogonal matrix $\mathbf{Q}$ and an $n \times n$ upper triangular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{QR}$.

same metrics. To solve this issue, we adopt the scheme of Yang *et al.* [21] by defining the following hybrid distance to measure the similarity between $\mathbf{x}_{\text{test}}$ and $\mathbf{x}_i$:

$$d(\mathbf{x}_{\text{test}}, \mathbf{x}_i) = (1-\mu)\frac{d_N(\mathbf{y}_{\text{test}}, \mathbf{y}_i)}{\sum_{j=1}^{n} d_N(\mathbf{y}_{\text{test}}, \mathbf{y}_j)} + \mu\frac{d_C(\mathbf{z}_{\text{test}}, \mathbf{z}_i)}{\sum_{j=1}^{n} d_C(\mathbf{z}_{\text{test}}, \mathbf{z}_j)} \quad (17)$$

where $\mu \in [0,1]$ is the fusion coefficient determining the weight of the two kinds of distances in the decision level.

Suppose that l^*_{test} is the class label of the test sample $\mathbf{x}_{\text{test}}$. Then, l^*_{test} can be obtained by

$$l^*_{\text{test}} = l_{i^*}, \qquad \text{where } i^* = \arg\min_i d(\mathbf{x}_{\text{test}}, \mathbf{x}_i). \quad (18)$$

IV. Extension to General Problems

In this section, we will generalize the solution method of the above KFSODVs/ROU algorithm to solve the more general cases like the following optimally constrained generalized Rayleigh quotient (OCGRQ) problems:

$$\begin{aligned}\omega_1 &= \arg\max_{\omega} \frac{\omega^T \mathbf{A}\omega}{\omega^T \mathbf{B}\omega}\\ &\cdots\\ \omega_k &= \arg \max_{\substack{\omega^T \mathbf{C}\omega_i = 0,\\ i=1,2,\ldots,k-1}} \frac{\omega^T \mathbf{A}\omega}{\omega^T \mathbf{B}\omega}\end{aligned} \quad (19)$$

where $\mathbf{A}$ and $\mathbf{C}$ are positive-semidefinite matrices, and $\mathbf{B}$ is a positive-definite matrix.[3] Similar with the solution method of KFSODVs/ROU, one can efficiently solve the OCGRQ problem via the ROU technique of the eigensystem, herein called OCGRQ/ROU.

Let $\mathbf{B} = \mathbf{U}_B\mathbf{\Lambda}_B\mathbf{U}_B^T$ be the SVD of $\mathbf{B}$ and let $\mathbf{L}_B = \mathbf{U}_B\mathbf{\Lambda}_B{}^{-1/2}$. Then, we have $\mathbf{L}_B^T\mathbf{B}\mathbf{L}_B = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Let $\omega = \mathbf{L}_B\alpha$. Then, solving (19) boils down to solving the following optimization problems:

$$\begin{aligned}\alpha_1 &= \arg \max_{\alpha^T\alpha=1} \alpha^T\mathbf{L}_B^T\mathbf{A}\mathbf{L}_B\alpha\\ &\cdots\\ \alpha_k &= \arg \max_{\substack{\alpha^T\alpha=1,\\ \alpha^T\mathbf{L}_B^T\mathbf{C}\mathbf{L}_B\alpha_i=0,\\ i=1,2,\ldots,k-1}} \alpha^T\mathbf{L}_B^T\mathbf{A}\mathbf{L}_B\alpha\\ &= \arg \max_{\substack{\alpha^T\alpha=1,\\ \alpha^T\mathbf{L}_B^T\mathbf{C}\mathbf{L}_B\mathbf{U}_{k-1}=\mathbf{0}}} \alpha^T\mathbf{L}_B^T\mathbf{A}\mathbf{L}_B\alpha\end{aligned} \quad (20)$$

where $\mathbf{U}_{k-1} = [\alpha_1, \alpha_2, \ldots, \alpha_{k-1}]$.

By comparing the formulation in (20) with that in (11), one can find that both equations can be unified into the same solution framework. If we simply replace the matrices $\mathbf{L}_w^T\mathbf{H}_b^{z^\perp}(\mathbf{H}_b^{z^\perp})^T\mathbf{L}_w$ and $\mathbf{\Lambda}_w^{-1}\mathbf{U}_{k-1}$ in (11) by $\mathbf{L}_B^T\mathbf{A}\mathbf{L}_B$ and $\mathbf{L}_B^T\mathbf{A}\mathbf{L}_B$, respectively, then one can see that the two equations are identical. Consequently, one can adopt the same fast algorithm of solving the discriminant vectors defined in (11) to solve the discriminant vectors of (20). We give the psudocode of solving the OCGRQ problem via the ROU technique (OCGRQ/ROU) in Algorithm 3.

[3]If $\mathbf{B}$ is singular, we can add a regularization term such that it is nonsingular; see Algorithm 3.

Algorithm 3: OCGRQ/ROU algorithm for solving the k optimal vectors $\omega_1, \ldots, \omega_k$ in (19)

Input:

- Data matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, the number k of optimal discriminant vectors, and the threshold $\varepsilon > 0$.

Initialization:

1) Perform SVD of $\mathbf{B}$: $\mathbf{B} = \mathbf{U}_B\mathbf{\Lambda}_B\mathbf{U}_B^T$. If $\mathbf{\Lambda}_B$ is singular, $\mathbf{\Lambda}_B \leftarrow \mathbf{\Lambda}_B + \text{diag}\{0,\ldots,0,\rho,\ldots,\rho\}$, ρ is the estimated noise spectrum [31].
2) $\mathbf{L}_B \leftarrow \mathbf{U}_B\mathbf{\Lambda}_B^{-1/2}$, $\tilde{\mathbf{C}} \leftarrow \mathbf{L}_B^T\mathbf{C}\mathbf{L}_B$, $\mathbf{A}_1 \leftarrow \mathbf{L}_B^T\mathbf{A}\mathbf{L}_B$.

For $i = 1, 2, \ldots, k$, **Do**

1) Solve the principal eigenvector α of $\mathbf{A}_i$ via the following power method:
   While $\|\alpha^{(j)} - \alpha^{(j-1)}\| > \varepsilon$, do
   - $\mathbf{z}^{(j)} = \mathbf{A}_i\alpha^{(j-1)}$ and $\alpha^{(j)} = \mathbf{z}^{(j)}/\|\mathbf{z}^{(j)}\|$;
   - $j = j + 1$.
2) $\omega_i \leftarrow \mathbf{L}_B\alpha$ and $\omega_i \leftarrow \omega_i/\|\omega_i\|$.
3) If $i = 1$, $\mathbf{q}_i \leftarrow \tilde{\mathbf{C}}\alpha$, $\mathbf{q}_i \leftarrow \mathbf{q}_i/\|\mathbf{q}_i\|$, and $\mathbf{Q}_1 \leftarrow \mathbf{q}_i$; else $\mathbf{q}_i \leftarrow \tilde{\mathbf{C}}\alpha - \mathbf{Q}_{i-1}(\mathbf{Q}_{i-1}^T(\tilde{\mathbf{C}}\alpha))$, $\mathbf{q}_i \leftarrow \mathbf{q}_i/\|\mathbf{q}_i\|$, and $\mathbf{Q}_i \leftarrow (\mathbf{Q}_{i-1} \quad \mathbf{q}_i)$.
4) $\mathbf{A}_{i+1} \leftarrow \mathbf{A}_i - \mathbf{A}_i\mathbf{q}_i\mathbf{q}_i^T - \mathbf{q}_i(\mathbf{q}_i^T\mathbf{A}_i) + (\mathbf{q}_i^T\mathbf{A}_i\mathbf{q}_i)\mathbf{q}_i\mathbf{q}_i^T$.

Output:

- $\omega_1, \omega_2, \ldots, \omega_k$.

The OCGRQ framework represents a large family of optimization problems in the dimensionality reduction field. Many current dimensionality reduction methods, such as ULDA [13], OLPP [14], and those methods in [15]–[19], can be unified into this framework. We will take the ULDA method as an example to show the application of the OCGRQ/ROU algorithm.

The ULDA method aims to find an optimal set of discriminant vectors that maximize Fisher's discriminant criterion $J(\omega) = \omega^T\mathbf{S}_b\omega/\omega^T\mathbf{S}_w\omega$ under the constraint $\omega_i^T\mathbf{S}_t\omega_j = 0$ $(i \neq j)$, where $\mathbf{S}_b$, $\mathbf{S}_w$, and $\mathbf{S}_t$ represent the between-class, within-class, and total-class scatter matrices, respectively. More specifically, the optimal set of discriminant vectors of ULDA can be formulated as the following optimization problems:

$$\begin{aligned}\omega_1 &= \arg \max_{\omega^T\omega=1} \frac{\omega^T\mathbf{S}_b\omega}{\omega^T\mathbf{S}_w\omega}\\ &\cdots\\ \omega_k &= \arg \max_{\substack{\omega^T\omega=1,\\ \omega^T\mathbf{S}_t\omega_i=0,\\ i=1,2,\ldots,k-1}} \frac{\omega^T\mathbf{S}_b\omega}{\omega^T\mathbf{S}_w\omega}.\end{aligned} \quad (21)$$

To solve the optimal discriminant vectors of ULDA, Jin *et al.* [13] proposed an algorithm based on the Lagrangian multipliers, herein called the ULDA/LM algorithm, which can be summarized as the following procedures.

1) The first optimal discriminant vector ω_1 is the eigenvector associated with the largest eigenvalue of the eigensystem

$$\mathbf{S}_w^{-1}\mathbf{S}_b\omega = \lambda\omega. \quad (22)$$

2) Suppose that the first r optimal discriminant vectors $\omega_1, \omega_2, \ldots, \omega_r$ have been obtained, then the $(r+1)$th optimal discriminant vector ω_{r+1} is the eigenvector associated with the largest eigenvalue of the eigensystem

$$\left(\mathbf{I}_d - \mathbf{S}_w^{-1}\mathbf{S}_t\mathbf{D}(\mathbf{D}^T\mathbf{S}_t\mathbf{S}_w^{-1}\mathbf{S}_t\mathbf{D})^{-1}\mathbf{D}^T\mathbf{S}_t\right) \times \mathbf{S}_w^{-1}\mathbf{S}_b\omega = \lambda\omega \quad (23)$$

where $\mathbf{D} = (\omega_1 \quad \omega_2 \quad \cdots \quad \omega_r)$ and $\mathbf{I}_d$ is the $d \times d$ identity matrix (we assume that the dimensionality of input data equals to d).

Similar to the KFSODVs/LM algorithm, solving each discriminant vector of ULDA using the ULDA/LM algorithm will involve computing the inverse of matrices, which will be costly when the number of k is large. However, if we adopt the OCGQR/ROU algorithm for solving ULDA, we can greatly reduce the computational cost. More specifically, let $\mathbf{A} = \mathbf{S}_b$, $\mathbf{B} = \mathbf{S}_w$, and $\mathbf{C} = \mathbf{S}_t$. Then, the solution of (21) can be obtained by using the OCGRQ/ROU algorithm. For simplicity, we call this new ULDA algorithm the ULDA/ROU algorithm.

V. Experiments

In this section, we will evaluate the performance of the proposed KFSODVs/ROU algorithms on six real data sets. For comparison, we conduct the same experiments using several commonly used kernel-based nonlinear feature extraction methods. These kernel-based methods include the KPCA method, the two-step kernel discriminant analysis (KDA) [22] method (i.e., KDA = KPCA + LDA [21]), and kernel direct discriminant analysis (KDDA) method [23]. We also conduct the same experiments using the orthogonal LDA (OLDA) method proposed by Ye [29] for further comparison. Although the discriminant vectors of OLDA are also orthogonal to each other, they are different from those of FSODVs because the discriminant vectors of OLDA are generated by performing the QR decomposition on the transformation matrix of ULDA [29]. So OLDA aims to find the orthogonal discriminant vectors from the subspace spanned by the discriminant vectors solved by ULDA, not the whole data space. Throughout the experiments, we use the nearest neighbor (NN) classifier [24] for the classification task. All the experiments are run on the platform of IBM personal computer with Matlab. The monomial kernel and the Gaussian kernel are, respectively, used to compute the elements of the Gram matrix $\mathbf{K} : (\mathbf{K})_{ij} = \ker(\mathbf{x}_i, \mathbf{x}_j)$ in the experiments, which are, respectively, defined as follows.

- Monomial kernel: $\ker(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y})^{dr}$, where dr is the monomial degree.
- Gaussian kernel: $\ker(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma)$, where σ is the parameter of the Gaussian kernel.

A. Brief Description of the Data Sets

The six data sets used in the experiments are: the United States Postal Service (USPS) handwritten digital database [12], the FERET database [25], the AR face database [26], the Olivetti Research Lab (ORL) face database in Cambridge,[4] the PIX face database from Manchester,[5] and the UMIST face database[6] [27]. The data sets are summarized as follows.

- The USPS database of handwritten digits is collected from mail envelopes in Buffalo, NY. This database consists of 7291 training samples and 2007 testing samples. Each sample is a 16×16 image denoting one of the ten digital characters {"0," "1," "2," "3," "4," "5," "6," "7," "8," "9"}.
- The FERET face database contains a total of 14 126 facial images by photographing 1199 subjects. All images are of size 512×768 pixels. We select a subset consisting of 984 images from 246 subjects who have "fa" images in both gallery "fa" and prob "duplicate" sets. For each subject, the four frontal images (two "fa" images plus two "fb" images) are used for our experiments, which involve variations of facial expression, aging, and wearing glasses. All images are manually cropped and normalized based on the distance between eyes and the distance between the eyes and the mouth and then are downsampled into a size of 72×64.
- The AR face database consists of over 3000 facial images of 126 subjects. Each subject has 26 facial images recorded in two different sessions, where each session consists of 13 images. The original image size is 768×576 pixels, and each pixel is represented by 24 b of RGB color values. We randomly select 70 subjects among the 126 subjects for the experiment. For each subject, only the 14 nonoccluded images are used for the experiment. All the selected images are centered and cropped to a size of 468×476 pixels, and then downsampled to a size of 25×25 pixels.
- The ORL face database contains 40 distinct subjects, where each one contains ten different images taken at different time with slightly varying lighting conditions. The original face images are all sized 92×112 pixels with a 256-level grayscale. The images are downsampled to a size of 23×28 in the experiment.
- The PIX face database contains 300 face images of 30 subjects. The original face images are all sized 512×512. We downsample each image to a size of 25×25 in the experiment.
- The UMIST face database contains 575 face images of 20 subjects. Each subject has a wide range of poses from profile to frontal views. We downsample each image to a size of 23×28 in the experiment.

B. Experiments on Testing the Performance of KFSODVs/ROU

In this experiment, we use the USPS database and the FERET database, respectively, to evaluate the performance of the KFSODVs/ROU algorithm in terms of the recognition accuracy and computational efficiency.

1) Experiments on the USPS Database: We choose the first 1000 training points of the USPS database to train the KFSODVs/ROU algorithm, and then use all the 2007 testing

[4]http://www.cam-orl.co.uk/facedatabase.html

[5]http://peipa.essex.ac.uk/ipa/pix/faces/manchester

[6]http://images.ee.umist.ac.uk/danny/database.html

TABLE III
RECOGNITION RATES (IN PERCENT) WITH DIFFERENT CHOICE OF THE MONOMIAL KERNEL DEGREE OF VARIOUS KERNEL-BASED FEATURE EXTRACTION METHODS ON THE USPS DATABASE. AS OLDA IS A LINEAR FEATURE EXTRACTION METHOD, IT DOES NOT HAVE HIGHER DEGREE COUNTERPARTS

Methods	Degree $dr = 1$	Degree $dr = 2$	Degree $dr = 3$	Degree $dr = 4$
KPCA	91.58	91.23	91.08	89.94
KPCA+LDA	86.55	92.03	92.33	91.98
KDDA	87.24	87.44	87.34	87.44
KFSODVs/ROU	91.38	92.53	92.97	92.78
OLDA	81.91	/	/	/

points to evaluate its recognition performance. The parameter μ is empirically fixed at 0.7 in the experiments. Table III shows the recognition rates of the various methods with different choice of the monomial kernel degrees. Note that OLDA is a linear feature extraction method. It can be regarded as using a monomial kernel with degree 1. From Table III, one can see that the best recognition rate is achieved by the KFSODVs/ROU method when the monomial kernel with degree 3 is used. One can also see from Table III that the FSODVs method (which is equivalent to the KFSODVS method when the monomial kernel with degree 1 is used) achieves much better performance than the OLDA method. This is simply because the orthogonal discriminant vectors of OLDA are obtained by performing the QR decomposition on the transformation matrix of ULDA [29], which contains only $c - 1 = 9$ columns in this experiment. Hence, the number of the discriminant vectors of OLDA is only 9. By comparison, however, the FSODVs method can obtain much more useful discriminant vectors and thus can obtain better recognition result.

To see the change of the recognition rates of the KFSODVs/ROU algorithm against the number of the discriminant vectors, we plot the recognition rates versus the number of the discriminant vectors in Fig. 1. It can be clearly seen from Fig. 1 that the best recognition rate of KFSODVs/ROU is achieved when the number of discriminant vectors is much larger than the number of the classes. For example, if the monomial kernel with degree 3 is used, the best recognition rate is achieved when the number of the discriminant vectors is larger than 300, much larger than the number of classes $(= 10)$. This testifies to the need of more than $c - 1$ discriminant vectors that we have claimed in the Introduction.

To compare the computational time spent on computing the discriminant vectors between the KFSODVs/ROU algorithm and the KFSODVs/LM algorithm, we plot in Fig. 2 the computation time with respect to the increase of the number of discriminant vectors to be computed for both algorithms, from which one can see that the computation time of KFSODVs/ROU is less than that of KFSODVs/LM, especially when the number of projection vectors is large. Moreover, it should be noted that the increase rate of the computation time of the KFSODVs/LM algorithm is faster than the KFSODVs/ROU algorithm with the increase of the number of projection vectors, where the increase of the central processing unit (CPU) time of the KFSODVs/LM method is nonlinear whereas the KFSODVs/ROU method is linear. The results shown in Fig. 2 coincide with our theoretical analysis in Section III. More specifically, we have theoretically shown in Section III that the computational complexity of solving the $(r+1)$th discriminant vector is $O(m^2)$ using the KFSODVs/ROU algorithm, where $m = \dim\{\overline{\mathbf{S}_w^y(0)}\}$ is a constant for a given problem. Consequently, the increase of the CPU time in computing discriminant vectors should be linear. Similar results can be seen in Figs. 4 and 5.

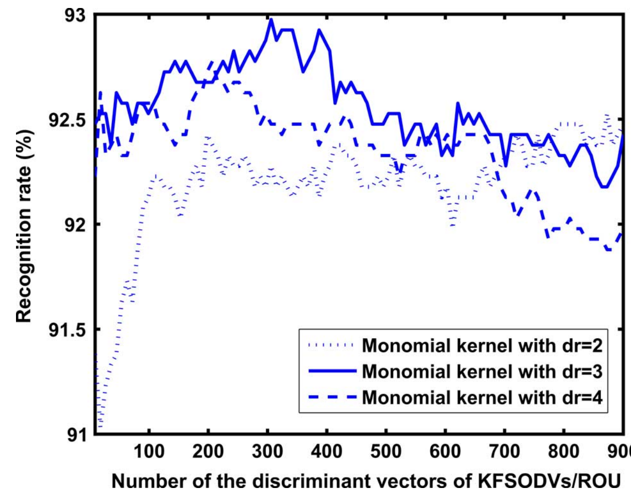

Fig. 1. Recognition rates versus the number of the discriminant vectors of KFSODVs/ROU on the USPS database.

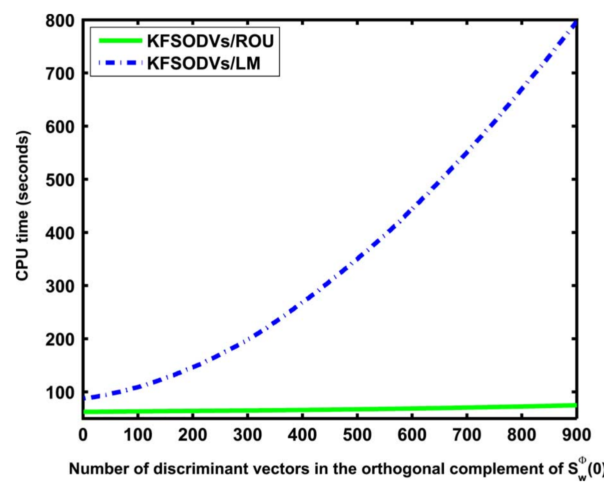

Fig. 2. Computational time with respect to the increase of the number of discriminant vectors for both the KFSODVs/ROU algorithm and the KFSODVs/LM algorithm on the USPS database.

2) Experiments on FERET Database: We use a fourfold cross-validation strategy to perform the experiment [28]: choose one image per subject as the testing sample and use the remaining three images per subject as the training samples. This procedure is repeated until all the images per subject have been used once as the testing sample. The final recognition rate is computed by averaging all the four trials. The parameter μ used in this experiment is fixed at 0.6.

Fig. 3 shows the recognition rates of the various kernel-based feature extraction methods as well as the OLDA method, where the monomial kernel with degree $dr = 1$ and the Gaussian kernel with parameter $\sigma = 2e8$ are, respectively, used to calculate the Gram matrix of the kernel-based algorithms. We can see from Fig. 3 that the KSFODVs/ROU method achieves much better results than the other methods.

We also conduct experiments to compare the computational time between the KFSODVs/ROU algorithm and the

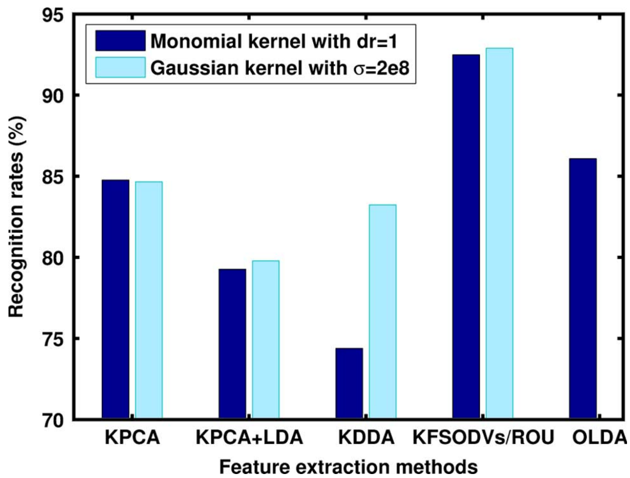

Fig. 3. Recognition rates of various kernel-based feature extraction methods on the FERET database.

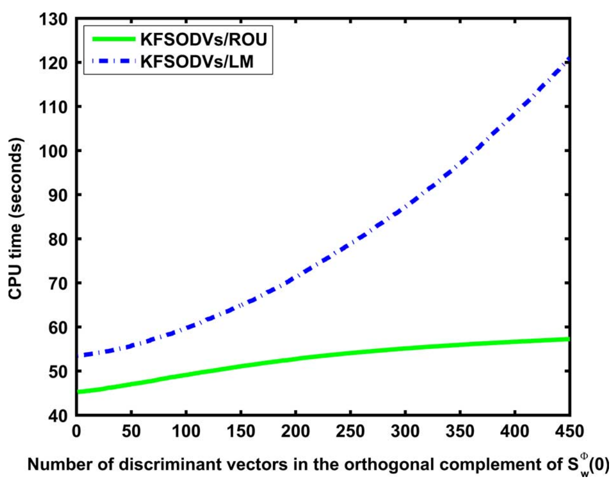

Fig. 4. Computation time with respect to the increase of the number of projection vectors for both the KFSODVs/ROU algorithm and the KFSODVs/LM algorithm on the FERET database.

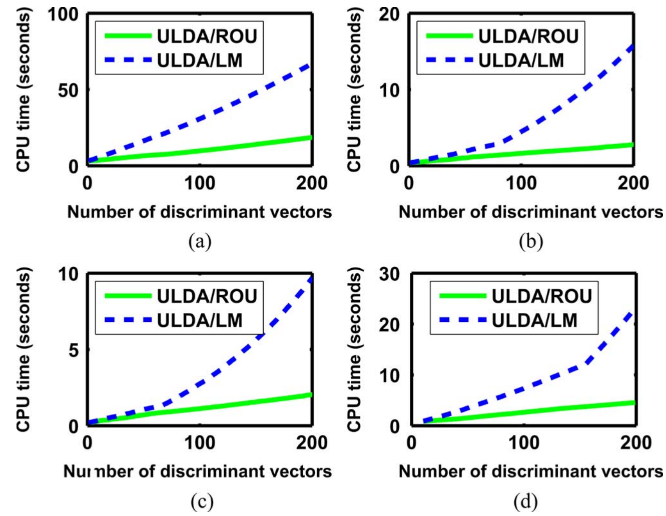

Fig. 5. Computation time with respect to the increase of the number of discriminant vectors for both the ULDA/ROU algorithm and the ULDA/LM algorithm. (a) AR database. (b) ORL database. (c) PIX database. (d) UMIST database.

KFSODVs/LM algorithm with respect to the increase of the number of discriminant vectors, where we use the training data of one trial as the experimental data and use the Gaussian kernel with parameter $\sigma = 2e8$ to calculate the Gram matrix. The experimental results are shown in Fig. 4, again from which one can see that the computation time of KFSODVs/ROU is less than that of KFSODVs/LM and as for the increasing rate of the computation time with the increase of the number of projection vectors, KFSODVs/LM is also faster than KFSODVs/ROU.

C. Experiments on Testing the Performance of ULDA/ROU

In this experiment, we aim to compare the computational efficiency between our ULDA/ROU algorithm and the ULDA/LM algorithm proposed by Jin *et al.* [13]. Four data sets, i.e., the AR database, the ORL database, the PIX database, and the UMIST database, are used as the experimental data to train both algorithms. The experiment is designed in the following form: we first randomly partition each data set into six subsets with approximately equal sizes, then we choose four of six subsets as the training data and the rest as the testing data. In the experiments of testing the recognition performance of ULDA/ROU, we conduct ten trials of experiments and get the average recognition rates as the final recognition rates. Fig. 5 shows the computation time of both algorithms with respect to the increase of the number of discriminant vectors to be computed. From Fig. 5, one can see that for the ULDA/LM algorithm, the increasing rate of the computation time with the increase of the number of projection vectors is faster than that of the ULDA/ROU algorithm.

VI. Conclusion and Discussion

In this paper, we have proposed a fast algorithm, called KFSODVs/ROU, to solve the kernel Foley–Sammon optimal discriminant vectors. Compared with the previous KFSODVs/LM algorithm, this new algorithm is more computationally efficient by using the MKGS algorithm to replace the KPCA algorithm as well as using the ROU technique of eigensystem to avoid the heavy cost of matrix computation. When solving each discriminant vector, the KFSODVs/ROU algorithm only needs a square complexity. However, the previous KFSODVs/LM algorithm needs a cubic complexity. The experimental results on both the USPS digital character database and the FERET face database have demonstrated the effectiveness of the proposed algorithm in terms of the computation efficiency and the recognition accuracy.

Although (K)FSODVs may not be needed when the number c of classes is comparable with the dimensionality of the feature vectors, (K)FSODVs are very helpful in improving the recognition rates when c is relatively small, as demonstrated by our experiments. Moreover, our algorithm is not limited to solving KFSODVs. It can be widely used for solving a family of OCGRQ problems with high efficiency. We take the ULDA method as an example and propose a new algorithm, called ULDA/ROU, for ULDA based on Algorithm 3. We also conduct extensive experiments on four face databases to show the computational efficiency of the ULDA/ROU algorithm with that of the previous ULDA/LM algorithm.

Additionally, other forms of the ROU technique may exist, resulting in other fast algorithms for solving KFSODVs. For example, the following formula may be used to compute the inverse of $\mathbf{D}_r^T(\mathbf{S}_w^{z^\perp})^{-1}\mathbf{D}_r$ recursively:

$$\begin{pmatrix}\mathbf{A} & \alpha\\ \beta^T & a\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{I}_r & -\mathbf{A}^{-1}\alpha\\ \mathbf{0}^T & 1\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1} & \mathbf{0}\\ \mathbf{0}^T & \frac{1}{a-\beta^T\mathbf{A}^{-1}\alpha}\end{pmatrix} \times \begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ -\beta^T\mathbf{A}^{-1} & 1\end{pmatrix}. \quad (24)$$

where $\mathbf{A}$ is an $r \times r$ nonsingular matrix, $\mathbf{I}_r$ is an $r \times r$ identity matrix, α and β are $r \times 1$ vectors, $\mathbf{0}$ represents the zero vectors, and c is a scalar. We will investigate such probabilities in the future work.

Appendix I

Proof of Lemma 1: We can find the complement basis $\mathbf{Q}^\perp$ such that the matrix $\tilde{\mathbf{Q}} = (\mathbf{Q} \quad \mathbf{Q}^\perp)$ is an orthogonal matrix. Then, we have $\mathbf{Q}^\perp(\mathbf{Q}^\perp)^T = \mathbf{I}_m - \mathbf{Q}\mathbf{Q}^T$ due to $\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^T = \mathbf{I}_m$. From $\alpha^T\mathbf{Q} = \mathbf{0}$, there exists a $\gamma \in \mathcal{R}^{m-r}$ such that

$$\alpha = \mathbf{Q}^\perp\gamma. \quad (25)$$

On the other hand, $\mathrm{rank}\{(\mathbf{Q}^\perp)^T\} = m - r$. Therefore, the columns of $(\mathbf{Q}^\perp)^T$ form a basis of $\mathcal{R}^{m-r}$. So there exists a $\beta \in \mathcal{R}^m$, such that

$$\gamma = (\mathbf{Q}^\perp)^T\beta. \quad (26)$$

Combining (25) and (26), we obtain

$$\alpha = \mathbf{Q}^\perp(\mathbf{Q}^\perp)^T\beta = (\mathbf{I}_m - \mathbf{Q}\mathbf{Q}^T)\beta. \quad (27)$$

Appendix II

Proof of Theorem 1: Since $\alpha^T\mathbf{\Lambda}_w^{-1}\mathbf{U}_r = \mathbf{0}$, $\mathbf{Q}_r\mathbf{R}_r$ is the QR decomposition of $\mathbf{\Lambda}_w^{-1}\mathbf{U}_r$, and $\mathbf{R}_r$ is nonsingular, we obtain that $\alpha^T\mathbf{Q}_r = \mathbf{0}$. From Lemma 1, there exists a $\beta \in \mathcal{R}^m$ such that $\alpha = (\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T)\beta$. Therefore

$$\begin{aligned}&\max_{\alpha^T\alpha=1,\alpha^T\mathbf{\Lambda}_w^{-1}\mathbf{U}_r=\mathbf{0}} \alpha^T\mathbf{L}_w^T\mathbf{H}_b^{z^\perp}\left(\mathbf{H}_b^{z^\perp}\right)^T\mathbf{L}_w\alpha\\ &= \max_\beta \frac{\beta^T(\mathbf{I}_m-\mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{L}_w^T\mathbf{H}_b^{z^\perp}(\mathbf{H}_b^{z^\perp})^T\mathbf{L}_w(\mathbf{I}_m-\mathbf{Q}_r\mathbf{Q}_r^T)\beta}{\beta^T(\mathbf{I}_m-\mathbf{Q}_r\mathbf{Q}_r^T)(\mathbf{I}_m-\mathbf{Q}_r\mathbf{Q}_r^T)\beta}\\ &= \max_\alpha \frac{\alpha^T(\mathbf{I}_m-\mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{L}_w^T\mathbf{H}_b^{z^\perp}(\mathbf{H}_b^{z^\perp})^T\mathbf{L}_w(\mathbf{I}_m-\mathbf{Q}_r\mathbf{Q}_r^T)\alpha}{\alpha^T\alpha}\end{aligned} \quad (28)$$

where we have utilized the idempotency of the matrix $\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T$ in the second equality. Therefore, α_{r+1} is the principal eigenvector corresponding to the largest eigenvalue of the matrix $(\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{L}_w^T\mathbf{H}_b^{z^\perp}{\mathbf{H}_b^{z^\perp}}^T\mathbf{L}_w(\mathbf{I}_m - \mathbf{Q}_r\mathbf{Q}_r^T)$.

Appendix III

Proof of Theorem 2: From $\mathbf{q} = \mathbf{\Lambda}_w^{-1}\alpha_{r+1} - \mathbf{Q}_r(\mathbf{Q}_r^T\mathbf{\Lambda}_w^{-1}\alpha_{r+1})$ and the fact that $\mathbf{Q}_r\mathbf{R}_r$ is the QR decomposition of $\mathbf{\Lambda}_w^{-1}\mathbf{U}_r$, we obtain that $\mathbf{Q}_r^T\mathbf{q} = \mathbf{0}$ and $\mathbf{Q}_r^T\mathbf{Q}_r = \mathbf{I}_r$, where $\mathbf{I}_r$ is the $r \times r$ identity matrix. Thus, we have

$$\begin{aligned}\mathbf{Q}_{r+1}^T\mathbf{Q}_{r+1} &= \begin{pmatrix}\mathbf{Q}_r & \frac{\mathbf{q}}{\|\mathbf{q}\|}\end{pmatrix}^T\begin{pmatrix}\mathbf{Q}_r & \frac{\mathbf{q}}{\|\mathbf{q}\|}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{Q}_r^T\mathbf{Q}_r & \mathbf{Q}_r^T\frac{\mathbf{q}}{\|\mathbf{q}\|}\\ \frac{\mathbf{q}^T\mathbf{Q}_r}{\|\mathbf{q}\|} & 1\end{pmatrix} = \mathbf{I}_{r+1}\end{aligned} \quad (29)$$

where $\mathbf{I}_{r+1}$ is the $(r+1) \times (r+1)$ identity matrix. On the other hand

$$\begin{aligned}\begin{pmatrix}\mathbf{Q}_r & \frac{\mathbf{q}}{\|\mathbf{q}\|}\end{pmatrix}\begin{pmatrix}\mathbf{R}_r & \mathbf{Q}_r^T\mathbf{\Lambda}_w^{-1}\alpha_{r+1}\\ \mathbf{0}^T & \|\mathbf{q}\|\end{pmatrix} &= (\mathbf{Q}_r\mathbf{R}_r \quad \mathbf{\Lambda}_w^{-1}\alpha_{r+1})\\ &= (\mathbf{\Lambda}_w^{-1}\mathbf{U}_r \quad \mathbf{\Lambda}_w^{-1}\alpha_{r+1})\\ &= \mathbf{\Lambda}_w^{-1}\mathbf{U}_{r+1}.\end{aligned} \quad (30)$$

From (29) and (30), one can see that the theorem is true.

Acknowledgment

The authors would like to thank one of the anonymous reviewers for pointing out the possibility of using the ROU technique for matrix inverses.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[3] P. Howland and H. Park, "Generalizing discriminant analysis using the generalized singular value decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 995–1006, Aug. 2004.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

[5] C. R. Rao, "The utilization of multiple measurements in problems of biological classification," *J. R. Statist. Soc. B*, vol. 10, pp. 159–203, 1948.

[6] 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.

[7] T. Okada and S. Tomita, "An optimal orthonormal system for discriminant analysis," *Pattern Recognit.*, vol. 18, no. 2, pp. 139–144, 1985.

[8] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 6, pp. 978–983, Jun. 1998.

[9] W. Zheng, L. Zhao, and C. Zou, "Foley-Sammon optimal discriminant vectors using kernel approach," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 1–9, Jan. 2005.

[10] W. Zheng, "Class-incremental generalized discriminant analysis," *Neural Comput.*, vol. 18, pp. 979–1006, 2006.

[11] W. Zheng, "Heteroscedastic feature extraction for texture classification," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 766–769, Sep. 2009.

[12] B. Schölkopf, A. J. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[13] 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.

[14] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal Laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006.

[15] K. Liu, Y.-Q. Cheng, and J.-Y. Yang, "A generalized optimal set of discriminant vectors," *Pattern Recognit.*, vol. 25, no. 7, pp. 731–739, 1992.

[16] A. K. Qin, P. N. Suganthan, and M. Loog, "Uncorrelated heteroscedastic LDA based on the weighted pairwise Chernoff criterion," *Pattern Recognit.*, vol. 38, pp. 613–616, 2005.
[17] Y. Liang, C. Li, W. Gong, and Y. Pan, "Uncorrelated linear discriminant analysis based on weighted pairwise Fisher criterion," *Pattern Recognit.*, vol. 40, pp. 3606–3615, 2007.
[18] Z. Liang and P. Shi, "Uncorrelated discriminant vectors using a kernel method," *Pattern Recognit.*, vol. 38, no. 2, pp. 307–310, 2005.
[19] D. Cai and X. He, "Orthogonal locality preserving indexing," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Session: Theory 1*, 2005, pp. 3–10.
[20] G. Golub and C. Van*, Matrix Computations*. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.
[21] J. Yang, A. F. Frangi, J.-Y. Yang, D. Zhang, and Z. Jin, "A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.
[22] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, pp. 2385–2404, 2000.
[23] 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.
[24] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[25] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[26] A. M. Martinez and R. Benavente, The AR Face Database, CVC Tech. Rep. #24, 1998.
[27] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures of general purpose face recognition," in *Face Recognition: From Theory to Application, NATO ASI Series F, Computer and Systems Sciences*, H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie, and T. S. Huang, Eds. New York: Springer-Verlag, 1998, vol. 163, pp. 446–456.
[28] K. Fukunaga*, Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.
[29] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *J. Mach. Learn. Res.*, vol. 6, pp. 483–502, 2005.
[30] X. Wang and X. Tang, "Unified subspace analysis for face recognition," in *Proc. 9th IEEE Conf. Comput. Vis.*, 2003, vol. 1, pp. 13–16.
[31] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," in *Proc. Int. Comput. Vis. Pattern Recognit.*, 2004, pp. 564–569.

Wenming Zheng (M'08) received the B.S. degree in computer science from Fuzhou University, Fuzhou, Fujian, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, Fujian, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, Jiangsu, China, in 2004.

Since 2004, he has been with the Research Center for Learning Science (RCLS), Southeast University, Nanjing, Jiangsu, China. His research interests include neural computation, pattern recognition, machine learning, and computer vision.

Zhouchen Lin (M'00–SM'08) received the B.S. degree in pure mathematics from Nankai University, Tianjin, China, in 1993, the M.S. degree in applied mathematics from Peking University, Beijing, China, in 1996, the M.Phil. degree in applies mathematics from Hong Kong Polytechnic University, Hong Kong, in 1997, and the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 2000.

Currently, he is a Researcher in Visual Computing Group, Microsoft Research Asia, Beijing, China. His research interests include computer vision, computer graphics, pattern recognition, statistical learning, document processing, and numerical computation and optimization.

Xiaoou Tang (S'93–M'96–SM'02–F'09) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1990, the M.S. degree from the University of Rochester, Rochester, NY, in 1991, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1996.

Currently, he is a Professor at the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. He worked as the Group Manager of the Visual Computing Group, Microsoft Research Asia, Beijing, China, from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing.

Dr. Tang received the Best Paper Award at the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). He is a Program Chair of the 2009 IEEE International Conference on Computer Vision (ICCV) and an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI) and the *International Journal of Computer Vision* (IJCV).