Discriminant Subspace Analysis: A Fukunaga-Koontz Approach

Sheng Zhang, *Member*, *IEEE*, and Terence Sim, *Member*, *IEEE*

Abstract—The Fisher Linear Discriminant (FLD) is commonly used in pattern recognition. It finds a linear subspace that maximally separates class patterns according to the Fisher Criterion. Several methods of computing the FLD have been proposed in the literature, most of which require the calculation of the so-called scatter matrices. In this paper, we bring a fresh perspective to FLD via the Fukunaga-Koontz Transform (FKT). We do this by decomposing the whole data space into four subspaces with different discriminabilities, as measured by eigenvalue ratios. By connecting the eigenvalue ratio with the generalized eigenvalue, we show where the Fisher Criterion is maximally satisfied. We prove the relationship between FLD and FKT analytically and propose a unified framework to understanding some existing work. Furthermore, we extend our theory to the Multiple Discriminant Analysis (MDA). This is done by transforming the data into intraclass and extraclass spaces, followed by maximizing the Bhattacharyya distance. Based on our FKT analysis, we identify the discriminant subspaces of MDA/FKT and propose an efficient algorithm, which works even when the scatter matrices are singular or too large to be formed. Our method is general and may be applied to different pattern recognition problems. We validate our method by experimenting on synthetic and real data.

Index Terms—Discriminant subspace analysis, Fukunaga-Koontz transform, pattern classification.

1 Introduction

In recent years, discriminant subspace analysis has been extensively studied in computer vision and pattern recognition. It has been widely used for feature extraction and dimensionality reduction in face recognition [2], [3], [15] and text classification [4]. One popular method is the Fisher Linear Discriminant (FLD), also known as the Linear Discriminant Analysis (LDA) [5], [7]. It tries to find an optimal subspace such that the separability of two classes is maximized. This is achieved by minimizing the within-class distance and maximizing the between-class distance simultaneously. To be more specific, in terms of the between-class scatter matrix $\mathbf{S}_b$ and the within-class scatter matrix $\mathbf{S}_w$, the Fisher Criterion can be written as

$$J_F(\Phi) = trace\Big\{(\Phi^T\mathbf{S}_w\Phi)^{-1}(\Phi^\top\mathbf{S}_b\Phi)\Big\}, \quad (1)$$

where Φ is a linear transformation matrix. By maximizing the criterion J_F, FLD finds the subspaces in which the classes are most linearly separable. The solution [7] that maximizes J_F is a set of the first eigenvectors $\{\phi_i\}$ that must satisfy

$$\mathbf{S}_b\phi = \lambda\mathbf{S}_w\phi. \quad (2)$$

This is called the generalized eigenvalue problem [5], [7]. The discriminant subspace is spanned by the generalized eigenvectors. The discriminability of each eigenvector is measured by the corresponding generalized eigenvalue, for example, the most discriminant subspace corresponds to the largest generalized eigenvalue. Equation (2) can be solved by matrix inversion and eigendecomposition, namely, by applying eigendecomposition on $\mathbf{S}_w^{-1}\mathbf{S}_b$. Unfortunately, for many applications with high-dimensional data and few training samples, for example, face recognition, the scatter matrix $\mathbf{S}_w$ is singular because, generally, the dimension of the data is larger than the number of samples. This is known as the undersampled or small sample size problem [7], [5].

Until now, many methods have been proposed to circumvent the requirement of nonsingularity of $\mathbf{S}_w$, such as Fisherface [2], Discriminant Common Vectors [3], Dual Space [19], LDA/GSVD [9], and LDA/QR [20]. In [2], Fisherface first applies PCA [13], [18] to reduce dimension such that $\mathbf{S}_w$ is nonsingular, then followed by LDA. The LDA/GSVD algorithm [9] avoids the inversion of $\mathbf{S}_w$ by the simultaneous diagonalization via GSVD. In [20], Ye and Li proposed a two-stage LDA method which applies QR decomposition on a small matrix, followed by LDA. Moreover, Ye and Li [20] also showed that both Fisherface and LDA/QR are the approximations of LDA/GSVD.

However, these methods do not directly relate to the generalized eigenvalue λ, the essential measure of discriminability. In fact, as we will show in Section 4, existing methods result in a suboptimum of the Fisher Criterion because important discriminant information is discarded to make $\mathbf{S}_w$ invertible. In our previous work [22], we proposed a better solution by applying the Fukunaga-Koontz Transform (FKT) to the LDA problem. Based on the eigenvalue ratio of FKT, we decomposed the whole data space into four subspaces. This revealed the relationship between LDA, FKT, and GSVD and allowed us to correctly maximize J_F even when $\mathbf{S}_w$ is singular.

In this paper, we extend our previous work in two ways: First, we present a unified framework for understanding other LDA-based methods. This provides valuable insights on how to choose the discriminant subspaces of the LDA problem. Second, we propose a new approach to multiple

- *S. Zhang is with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106-9560. E-mail: zhangs@ece.ucsb.edu.*
- *T. Sim is with the School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543. E-mail: tsim@comp.nus.edu.sg.*

Manuscript received 13 Feb. 2006; revised 28 Aug. 2006; accepted 13 Dec. 2006; published online 18 Jan. 2007.
Recommended for acceptance by M. Figueiredo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0150-0206.
Digital Object Identifier no. 10.1109/TPAMI.2007.1089.

discriminant analysis (MDA). This is done by casting the multiclass problem into a two-class one and by maximizing the Bhattacharyya distance (which is the error bound of the Bayes Classifier [5]) rather than the Fisher Criterion. Then, the discriminant subspace is obtained algebraically via FKT. This means that our method can find the global optimum directly (no iteration required), which is not the case in [6]. For completeness, in this paper, we include details of our previous work [22] as well.

To summarize, our work has three main contributions:

1. We present a unifying framework to understand different methods, namely, LDA, FKT, and GSVD. To be more specific, we show that, for the LDA problem, GSVD is equivalent to FKT and the generalized eigenvalue of LDA is equal to both the eigenvalue ratio of FKT and the square of the generalized singular value of GSVD.
2. We prove that our approach is useful for general pattern recognition. Our theoretical analyses demonstrate how to choose the best subspaces for maximum discriminability and unify other subspace methods such as Fisherface, PCA+NULL space, LDA/QR, and LDA/GSVD.
3. We further propose a new criterion to handle MDA, derived from the Bhattacharyya distance. Because the Bhattacharyya distance upper bounds the Bayes error [5], this new criterion is theoretically superior to the Fisher Criterion, which is not related to the Bayes error in general.

The rest of this paper is organized as follows: Section 2 reviews related work, that is, PCA, LDA, Fisherface, PCA+NULL Space, LDA/QR, and LDA/GSVD. We discuss FKT in Section 3, where discriminant subspace analysis based on FKT is also presented. In Section 4, we show how to unify some LDA-based methods based on our theory. Moreover, we demonstrate how to handle the multiclass problem by FKT in Section 5. We apply our theory to the classification problem on synthetic and real data in Section 6 and conclude our paper in Section 7.

2 Related Work

Notation. Let $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$, $\mathbf{a}_i \in \mathbb{R}^D$ denote a data set of given D-dimensional vectors. Each data point belongs to exactly one of C object classes $\{L_1, \ldots, L_C\}$. The number of vectors in class L_i is denoted by N_i; thus, $N = \sum N_i$. Observe that, for high-dimensional data, for example, face images, generally, $C \leq N \ll D$. The between-class scatter matrix $\mathbf{S}_b$, the within-class scatter matrix $\mathbf{S}_w$, and the total scatter matrix $\mathbf{S}_t$ are defined as follows:

$$\mathbf{S}_b = \sum_{i=1}^{C} N_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^\top = \mathbf{H}_b\mathbf{H}_b^\top, \tag{3}$$

$$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{\mathbf{a}_j \in L_i} (\mathbf{a}_j - \mathbf{m}_i)(\mathbf{a}_j - \mathbf{m}_i)^\top = \mathbf{H}_w\mathbf{H}_w^\top, \tag{4}$$

$$\mathbf{S}_t = \sum_{i=1}^{N} (\mathbf{a}_i - \mathbf{m})(\mathbf{a}_i - \mathbf{m})^\top = \mathbf{H}_t\mathbf{H}_t^\top, \tag{5}$$

$$\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w. \tag{6}$$

Here, $\mathbf{m}_i$ denotes the class mean and $\mathbf{m}$ is the global mean of $\mathbf{A}$. The matrices $\mathbf{H}_b \in \mathbb{R}^{D\times C}$, $\mathbf{H}_w \in \mathbb{R}^{D\times N}$, and $\mathbf{H}_t \in \mathbb{R}^{D\times N}$ are the precursor matrices of the between-class scatter matrix, the within-class scatter matrix, and the total scatter matrix, respectively,

$$\mathbf{H}_b = \left[\sqrt{N_1}(\mathbf{m}_1 - \mathbf{m}), \ldots, \sqrt{N_c}(\mathbf{m}_C - \mathbf{m})\right], \tag{7}$$

$$\mathbf{H}_w = \left[\mathbf{A}_1 - \mathbf{m}_1 \cdot \mathbf{1}_1^\top, \ldots, \mathbf{A}_C - \mathbf{m}_C \cdot \mathbf{1}_C^\top\right], \tag{8}$$

$$\mathbf{H}_t = [\mathbf{a}_1 - \mathbf{m}, \ldots, \mathbf{a}_N - \mathbf{m}]. \tag{9}$$

Here, $\mathbf{1}_i = (1, \ldots, 1)^\top \in \mathbb{R}^{N_i}$ and $\mathbf{A}_i$ is the data matrix for class L_i. Let us denote the rank of each scatter matrix by $r_w = rank(\mathbf{S}_w)$, $r_b = rank(\mathbf{S}_b)$, and $r_t = rank(\mathbf{S}_t)$. Note that, for high-dimensional data ($N \ll D$), $r_b \leq C - 1$, $r_w \leq N - C$, and $r_t \leq N - 1$.

2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [13] is one of the well-known subspace methods for dimensionality reduction. It is the optimal method for statistical pattern representation in terms of the mean square error. PCA can be readily computed by applying the eigendecomposition on the total scatter matrix, that is, $\mathbf{S}_t = \mathbf{U}\mathbf{D}\mathbf{U}^\top$. By keeping the eigenvectors (principal components) corresponding to the largest eigenvalues, we can compute the PCA projection matrix. To solve the appearance-based face recognition problem, Turk and Pentland [18] proposed "Eigenface" by using PCA. Note that PCA is optimal for pattern representation, not necessarily for classification [5]. LDA [5], however, is another well-known subspace method designed for pattern classification.

2.2 Linear Discriminant Analysis (LDA)

Given the data matrix $\mathbf{A}$, which can be divided into C classes, we try to find a linear transformation matrix $\Phi \in \mathbb{R}^{D\times d}$, where $d < D$. This maps a high-dimensional data to a low-dimensional space. From the perspective of pattern classification, LDA aims to find the optimal transformation Φ such that the projected data are well separated.

Regarding pattern classification, usually, there are two types of criteria that are used to measure the separability of classes [7]. One is a family of criteria that gives the upper bound on the Bayes error, for example, Bhattacharyya distance. The other is based on a family of functions of scatter matrices. As shown in (1), the Fisher Criterion belongs to the latter one. Moreover, the solution of the criterion is the generalized eigenvector and eigenvalue of the scatter matrices (see (2)). However, if $\mathbf{S}_w$ is nonsingular, it can be solved by the generalized eigendecomposition: $\mathbf{S}_w^{-1}\mathbf{S}_b\phi = \lambda\phi$. Otherwise, $\mathbf{S}_w$ is singular and we circumvent this by methods such as Fisherface [2], PCA+NULL Space [10], LDA/QR [20], and LDA/GSVD [9].

2.3 Fisherface

To handle face recognition under different lightings, Belhumeur et al. [2] proposed "Fisherface," which is an application of LDA. In the Fisherface method, PCA is performed first so as to make $\mathbf{S}_w$ nonsingular, followed by LDA. This means that $\text{Fisherface} = \text{LDA} + \text{PCA}$. However, there exist at least two problems: 1) During PCA, it is not clear how many dimensions should be kept so that $\mathbf{S}_w$ is nonsingular and 2) to avoid the singularity of $\mathbf{S}_w$, some

directions/eigenvectors (corresponding to the small nonzero eigenvalues) are thrown away in the PCA step, which may contain discriminant information [21].

2.4 PCA + NULL Space

Considering that the null space of $\mathbf{S}_w$ contains discriminant information, Huang et al. [10] first remove the null space of $\mathbf{S}_t$. This is the intersection of null space of $\mathbf{S}_b$ and $\mathbf{S}_w$ and has been proven to be useless for discrimination [10]. It can be done by applying PCA first, followed by computing the principal components of $\mathbf{S}_b$ within the null space of $\mathbf{S}_w$. More precisely, it is realized in three steps:

- **Step 1**. Remove the null space of $\mathbf{S}_t$: Eigendecompose $\mathbf{S}_t$, $\mathbf{S}_t = \mathbf{U}\mathbf{D}_t\mathbf{U}^\top$, and $\mathbf{U}$ is the set of eigenvectors corresponding to the nonzero eigenvalues. Let $\mathbf{S}'_w = \mathbf{U}^\top\mathbf{S}_w\mathbf{U}$ and $\mathbf{S}'_b = \mathbf{U}^\top\mathbf{S}_b\mathbf{U}$.
- **Step 2**. Compute the null space of $\mathbf{S}'_w$: Eigendecompose $\mathbf{S}'_w$ and let $\mathbf{Q}_\perp$ be the set of eigenvectors corresponding to the zero eigenvalues. Let $\mathbf{S}''_b = \mathbf{Q}_\perp^\top\mathbf{S}'_b\mathbf{Q}_\perp$.
- **Step 3**. Remove the null space of $\mathbf{S}''_b$ if it exists: Eigendecompose $\mathbf{S}''_b$ and keep the set of eigenvectors corresponding to the nonzero eigenvalues.

The key difference between PCA+NULL Space and Fisherface is in the first step: PCA+NULL Space only removes the eigenvectors with zero eigenvalues, whereas Fisherface removes eigenvectors corresponding to zero and nonzero eigenvalues.

2.5 LDA/QR

In [20], Ye and Li proposed a two-stage LDA method, namely, LDA/QR. It not only overcomes the singularity problems of LDA but also achieves computational efficiency. This is done by applying QR decomposition on $\mathbf{H}_b$ first, followed by LDA. To be more specific, it is realized in two steps:

- **Step 1**. Apply QR decomposition on $\mathbf{H}_b$: $\mathbf{H}_b = \mathbf{Q}\mathbf{R}$, where $\mathbf{Q} \in \mathbb{R}^{D\times r_b}$ has orthogonal columns that span the space of $\mathbf{H}_b$ and $\mathbf{R} \in \mathbb{R}^{r_b\times C}$ is an upper triangular matrix. Then, define $\widetilde{\mathbf{S}}_b = \mathbf{Q}^\top\mathbf{S}_b\mathbf{Q}$ and $\widetilde{\mathbf{S}}_w = \mathbf{Q}^\top\mathbf{S}_w\mathbf{Q}$.
- **Step 2**. Apply LDA on $\widetilde{\mathbf{S}}_b$ and $\widetilde{\mathbf{S}}_w$: Keep the set of eigenvectors corresponding to the smallest eigenvalues of $\widetilde{\mathbf{S}}_b^{-1}\widetilde{\mathbf{S}}_w$.

Note that, to reduce computational load, QR decomposition is employed here, whereas, in the Fisherface and PCA+NULL space methods, the subspace is obtained by using eigendecomposition.

2.6 LDA/GSVD

The GSVD was originally defined by Van Loan [14] and then Page and Saunders [16] extended it to handle any two matrices with the same number of columns. We will briefly review the mechanism of GSVD, using LDA as an example.

Howland and Park [9] extended the applicability of LDA to the case when $\mathbf{S}_w$ is singular. This is done by using simultaneous diagonalization of the scatter matrices via the GSVD [8]. First, to reduce computational load, $\mathbf{H}_b$ and $\mathbf{H}_w$ are used instead of $\mathbf{S}_b$ and $\mathbf{S}_w$. Then, based on GSVD, there exist orthogonal matrices $\mathbf{Y} \in \mathbb{R}^{C\times C}$ and $\mathbf{Z} \in \mathbb{R}^{N\times N}$ and a nonsingular matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$ such that

$$\mathbf{Y}^\top\mathbf{H}_b^\top\mathbf{X} = [\Sigma_b, \mathbf{0}], \tag{10}$$

$$\mathbf{Z}^\top\mathbf{H}_w^\top\mathbf{X} = [\Sigma_w, \mathbf{0}], \tag{11}$$

where

$$\Sigma_b = \begin{bmatrix} \mathbf{I}_b & & \\ & \mathbf{D}_b & \\ & & \mathbf{O}_b \end{bmatrix}, \qquad \Sigma_w = \begin{bmatrix} \mathbf{O}_w & & \\ & \mathbf{D}_w & \\ & & \mathbf{I}_w \end{bmatrix}.$$

The matrices $\mathbf{I}_b \in \mathbb{R}^{(r_t-r_w)\times(r_t-r_w)}$ and $\mathbf{I}_w \in \mathbb{R}^{(r_t-r_b)\times(r_t-r_b)}$ are identity matrices, $\mathbf{O}_b \in \mathbb{R}^{(C-r_b)\times(r_t-r_b)}$ and $\mathbf{O}_w \in \mathbb{R}^{(N-r_w)\times(r_t-r_w)}$ are rectangular zero matrices that may have no rows or no columns, $\mathbf{D}_b = \mathrm{diag}(\alpha_{r_t-r_w+1}, \ldots, \alpha_{r_b})$ and $\mathbf{D}_w = \mathrm{diag}(\beta_{r_t-r_w+1}, \ldots, \beta_{r_b})$ satisfy $1 > \alpha_{r_t-r_w+1} \geq \ldots \geq \alpha_{r_b} > 0$, $0 < \beta_{r_t-r_w+1} \leq \ldots \leq \beta_{r_b} < 1$, and $\alpha_i^2 + \beta_i^2 = 1$. Thus, $\Sigma_b^\top\Sigma_b + \Sigma_w^\top\Sigma_w = \mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{r_t\times r_t}$ is an identity matrix. The columns of $\mathbf{X}$, which are the generalized singular vectors for the matrix pair $[\mathbf{H}_b, \mathbf{H}_w]$, can be used as the discriminant feature subspace based on GSVD.

3 Fukunaga-Koontz Transform and LDA

In this section, we begin by briefly reviewing the Fukunaga-Koontz Transform (FKT). Then, based on the eigenvalue ratio of FKT, we analyze the discriminant subspaces by breaking the whole space into smaller subspaces. Finally, we connect FKT to the Fisher Criterion, which suggests a way to select discriminant subspaces.

3.1 Fukunaga-Koontz Transform

The FKT was designed for the two-class recognition problem. Given the data matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ from two classes, the autocorrelation matrices $\mathbf{S}_1 = \mathbf{A}_1\mathbf{A}_1^\top$ and $\mathbf{S}_2 = \mathbf{A}_2\mathbf{A}_2^\top$ are positive semidefinite (p.s.d.) and symmetric. The sum of these two matrices is still p.s.d. and symmetric and can be factorized in the form

$$\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2 = [\mathbf{U}, \mathbf{U}_\perp]\begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{U}^\top \\ \mathbf{U}_\perp^\top \end{bmatrix}. \tag{12}$$

Without loss of generality, $\mathbf{S}$ may be singular and $r = rank(\mathbf{S}) < D$; thus, $\mathbf{D} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_r\}$, $\lambda_1 \geq \ldots \geq \lambda_r > 0$. The set of eigenvectors $\mathbf{U} \in \mathbb{R}^{D\times r}$ corresponds to nonzero eigenvalues and the set $\mathbf{U}_\perp \in \mathbb{R}^{D\times(D-r)}$ is the orthogonal complement of $\mathbf{U}$. Now, we can whiten $\mathbf{S}$ by a transformation operator $\mathbf{P} = \mathbf{U}\mathbf{D}^{-1/2}$. The sum of the two matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ becomes

$$\mathbf{P}^\top\mathbf{S}\mathbf{P} = \mathbf{P}^\top(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{P} = \widetilde{\mathbf{S}}_1 + \widetilde{\mathbf{S}}_2 = \mathbf{I}, \tag{13}$$

where $\widetilde{\mathbf{S}}_1 = \mathbf{P}^\top\mathbf{S}_1\mathbf{P}$, $\widetilde{\mathbf{S}}_2 = \mathbf{P}^\top\mathbf{S}_2\mathbf{P}$, and $\mathbf{I} \in \mathbb{R}^{r\times r}$ is an identity matrix. Suppose the eigenvector of $\widetilde{\mathbf{S}}_1$ is $\mathbf{v}$ with the eigenvalue λ_1, that is, $\widetilde{\mathbf{S}}_1\mathbf{v} = \lambda_1\mathbf{v}$. Since $\widetilde{\mathbf{S}}_1 = \mathbf{I} - \widetilde{\mathbf{S}}_2$, we can rewrite it as

$$(\mathbf{I} - \widetilde{\mathbf{S}}_2)\mathbf{v} = \lambda_1\mathbf{v}, \tag{14}$$

$$\widetilde{\mathbf{S}}_2\mathbf{v} = (1 - \lambda_1)\mathbf{v}. \tag{15}$$

This means that $\widetilde{\mathbf{S}}_2$ has the same eigenvector as $\widetilde{\mathbf{S}}_1$, but the corresponding eigenvalue is $\lambda_2 = 1 - \lambda_1$. Consequently, the dominant eigenvector of $\widetilde{\mathbf{S}}_1$ is the weakest eigenvector of $\widetilde{\mathbf{S}}_2$ and vice versa. This suggests that a pattern belonging to Class 1 ought to yield a large coefficient when projected onto

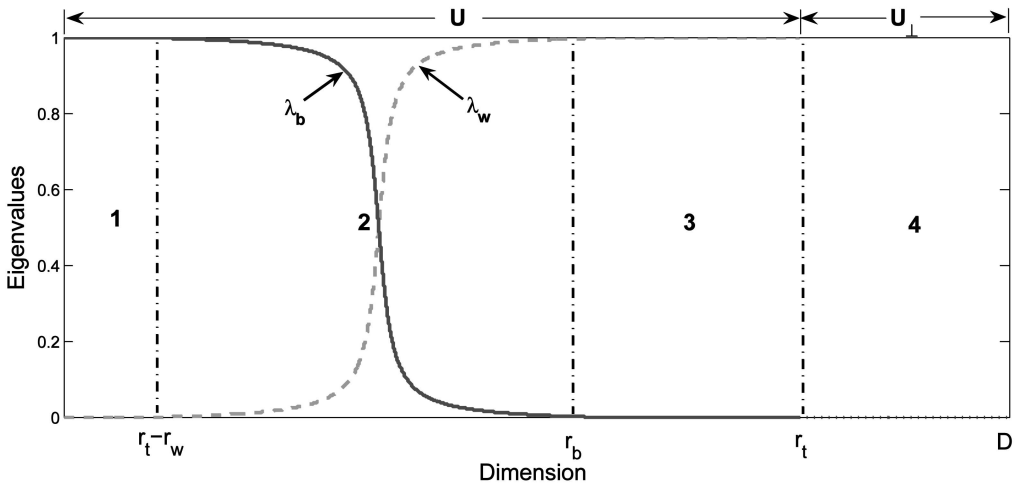


Fig. 1. The whole data space is decomposed into four subspaces via FKT. In $\mathbf{U}_\perp$, the null space of $\mathbf{S}_t$, there is no discriminant information. In $\mathbf{U}$, $\lambda_b + \lambda_w = 1$. Note that we represent all possible subspaces, but, in real cases, some of these subspaces may not be available.

the dominant eigenvector of $\widetilde{\mathbf{S}}_1$ and vice versa. The dominant eigenvectors therefore form a subspace in which the two classes are separable. Classification can then be done by, say, picking the nearest neighbor (NN) in this subspace.

Recently, it was proven that, under certain conditions, FKT is the best linear approximation to a quadratic classifier [11]. Interested readers may refer to [7] and [11] for more details.

3.2 LDA/FKT

Generally speaking, for the LDA problem, there are more than two classes. To handle the multiclass problem, we replace the autocorrelation matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ with the scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$. Since $\mathbf{S}_b$, $\mathbf{S}_w$, and $\mathbf{S}_t$ are p.s.d. and symmetric and $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$, we can apply FKT on $\mathbf{S}_b$, $\mathbf{S}_w$, and $\mathbf{S}_t$, which is called LDA/FKT hereafter in this paper. The whole data space is decomposed into $\mathbf{U}$ and $\mathbf{U}_\perp$ (Fig. 1). On one hand, $\mathbf{U}_\perp$ is the set of eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_t$. This has been proven to be the intersection of the null spaces of $\mathbf{S}_b$ and $\mathbf{S}_w$ and contains no discriminant information [10]. On the other hand, $\mathbf{U}$ is the set of eigenvectors corresponding to the nonzero eigenvalues of $\mathbf{S}_t$. It contains discriminant information.

Based on FKT, $\widetilde{\mathbf{S}}_b = \mathbf{P}^\top \mathbf{S}_b \mathbf{P}$ and $\widetilde{\mathbf{S}}_w = \mathbf{P}^\top \mathbf{S}_w \mathbf{P}$ share the same eigenspace and the sum of two eigenvalues corresponding to the same eigenvector is equal to 1.

$$\widetilde{\mathbf{S}}_b = \mathbf{V}\Lambda_b\mathbf{V}^\top, \tag{16}$$

$$\widetilde{\mathbf{S}}_w = \mathbf{V}\Lambda_w\mathbf{V}^\top, \tag{17}$$

$$\mathbf{I} = \Lambda_b + \Lambda_w. \tag{18}$$

Here, $\mathbf{V} \in \mathbb{R}^{r_t \times r_t}$ is the orthogonal eigenvector matrix and Λ_b, $\Lambda_w \in \mathbb{R}^{r_t \times r_t}$ are diagonal eigenvalue matrices. According to the eigenvalue ratio $\frac{\lambda_b}{\lambda_w}$, $\mathbf{U}$ can be further decomposed into three subspaces. To keep the integrity of the whole data space, we incorporate $\mathbf{U}_\perp$ as the fourth subspace (Fig. 1):

1. **Subspace 1**. $\mathrm{span}(\mathbf{S}_b) \bigcap \mathrm{null}(\mathbf{S}_w)$, the set of eigenvectors $\{\mathbf{v}_i\}$ corresponding to $\lambda_w = 0$ and $\lambda_b = 1$. Since $\frac{\lambda_b}{\lambda_w} = \infty$, in this subspace, the eigenvalue ratio is maximized.
2. **Subspace 2**. $\mathrm{span}(\mathbf{S}_b) \bigcap \mathrm{span}(\mathbf{S}_w)$, the set of eigenvectors $\{\mathbf{v}_i\}$ corresponding to $0 < \lambda_w < 1$ and $0 < \lambda_b < 1$. Since $0 < \frac{\lambda_b}{\lambda_w} < \infty$, the eigenvalue ratio is finite and smaller than that of Subspace 1.
3. **Subspace 3**. $\mathrm{null}(\mathbf{S}_b) \bigcap \mathrm{span}(\mathbf{S}_w)$, the set of eigenvectors $\{\mathbf{v}_i\}$ corresponding to $\lambda_w = 1$ and $\lambda_b = 0$. Since $\frac{\lambda_b}{\lambda_w} = 0$, the eigenvalue ratio is minimum.
4. **Subspace 4**. $\mathrm{null}(\mathbf{S}_b) \bigcap \mathrm{null}(\mathbf{S}_w)$, the set of eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_t$.

Note that, in practice, some of these four subspaces may not exist, depending on the ranks of $\mathbf{S}_b$, $\mathbf{S}_w$, and $\mathbf{S}_t$. As illustrated in Fig. 1, the null space of $\mathbf{S}_w$ is the union of Subspace 1 and Subspace 4, whereas the null space of $\mathbf{S}_b$ is the union of Subspace 3 and Subspace 4, if they exist. Therefore, from the perspective of FKT, we reach the same conclusion as Huang et al. in [10]. That is, Subspace 4 is the intersection of the null spaces of $\mathbf{S}_b$ and $\mathbf{S}_w$.

3.3 Relationship between FKT, GSVD, and LDA

How do these four subspaces help to maximize the Fisher Criterion J_F? We explain this in Theorem 1, which connects the generalized eigenvalue of J_F to the eigenvalues of FKT. We begin with a lemma (see Appendix A for the proof):

Lemma 1. *For the LDA problem, GSVD is equivalent to FKT, with* $\mathbf{X} = [\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp]$, $\Lambda_b = \Sigma_b^\top \Sigma_b$, *and* $\Lambda_w = \Sigma_w^\top \Sigma_w$, *where* $\mathbf{X}$, Σ_b, *and* Σ_w *are from GSVD (10), (11), and* $\mathbf{U}$, $\mathbf{D}$, $\mathbf{V}$, $\mathbf{U}_\perp$, Λ_w, *and* Λ_b *are matrices from FKT (16), (17), (18).*

Now, based on the above lemma, we can investigate the relationship between the eigenvalue ratio of FKT and the generalized eigenvalue λ of the Fisher Criterion J_F.

Theorem 1. *If* λ *is the solution of (2) (the generalized eigenvalue of* $\mathbf{S}_b$ *and* $\mathbf{S}_w$*) and* λ_b *and* λ_w *are the eigenvalues after applying FKT on* $\mathbf{S}_b$ *and* $\mathbf{S}_w$, *then* $\lambda = \frac{\lambda_b}{\lambda_w}$, *where* $\lambda_b + \lambda_w = 1$.

Proof. Based on GSVD, it is easy to verify that

$$\mathbf{S}_b = \mathbf{H}_b\mathbf{H}_b^\top = \mathbf{X}^{-\top}\begin{bmatrix} \Sigma_b^\top\Sigma_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{X}^{-1}. \tag{19}$$

TABLE 1
Comparison between Different Methods: N Is the Number of Training Samples, D Is the Dimension, and C Is the Number of Classes

Method	Time complexity	Space complexity	Discriminant subspaces	Remarks
PCA	$O(DN^2)$	$O(DN)$	N/A	Optimal for pattern representation
Fisherface	$O(DN^2)$	$O(DN)$	2 and 3	Sub-optimal for the Fisher Criterion
PCA+NULL	$O(DN^2)$	$O(DN)$	1	
LDA/QR	$O(DNC)$	$O(DC)$	1 and 2	
LDA/GSVD	$O((N+C)^2D)$	$O(DN)$	1, 2, 3 and 4	Optimal for the Fisher Criterion
LDA/FKT	$O(DN^2)$	$O(DN)$	1, 2 and 3	
MDA/FKT	$O(DN^2)$	$O(DN)$	1, 2 and 3	Optimal for Bhattacharyya distance

For discriminant subspaces, please refer to Fig. 1.

According to Lemma 1, $\Lambda_b = \Sigma_b^\top \Sigma_b$, thus,

$$\mathbf{S}_b = \mathbf{X}^{-\top} \begin{bmatrix} \Lambda_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{X}^{-1}. \tag{20}$$

Similarly,

$$\mathbf{S}_w = \mathbf{X}^{-\top} \begin{bmatrix} \Lambda_w & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{X}^{-1}. \tag{21}$$

Since $\mathbf{S}_b\phi = \lambda \mathbf{S}_w\phi$,

$$\mathbf{X}^{-\top} \begin{bmatrix} \Lambda_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{X}^{-1}\phi = \lambda \mathbf{X}^{-\top} \begin{bmatrix} \Lambda_w & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{X}^{-1}\phi. \tag{22}$$

Letting $\mathbf{v} = \mathbf{X}^{-1}\phi$, and multiplying $\mathbf{X}^\top$ on both sides, we obtain the following:

$$\begin{bmatrix} \Lambda_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{v} = \lambda \begin{bmatrix} \Lambda_w & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{v}. \tag{23}$$

If we add $\lambda \begin{bmatrix} \Lambda_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ on both sides of the above equation, then

$$(1+\lambda) \begin{bmatrix} \Lambda_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{v} = \lambda \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{v}. \tag{24}$$

This means that $(1+\lambda)\lambda_b = \lambda$, which can be rewritten as $\lambda_b = \lambda(1-\lambda_b) = \lambda\lambda_w$ because $\lambda_b + \lambda_w = 1$. Now, we can observe that $\lambda = \frac{\lambda_b}{\lambda_w}$. □

Corollary 1. *If λ is the generalized eigenvalue of $\mathbf{S}_b$ and $\mathbf{S}_w$, α and β are the solutions of (10) and (11), and α/β is the generalized singular value of the matrix pair $(\mathbf{H}_b, \mathbf{H}_w)$, then $\lambda = \frac{\alpha^2}{\beta^2}$, where $\alpha^2 + \beta^2 = 1$.*

Proof. In Lemma 1, we have proven that $\Lambda_b = \Sigma_b^\top \Sigma_b$ and $\Lambda_w = \Sigma_w^\top \Sigma_w$, that is, $\lambda_b = \alpha^2$ and $\lambda_w = \beta^2$. According to Theorem 1, we observe that $\lambda = \frac{\lambda_b}{\lambda_w}$. Therefore, it is easy to see that $\lambda = \frac{\alpha^2}{\beta^2}$. Note that $\frac{\alpha}{\beta}$ is the generalized singular value of $(\mathbf{H}_b, \mathbf{H}_w)$ by GSVD and λ is the generalized eigenvalue of $(\mathbf{S}_b, \mathbf{S}_w)$. □

The corollary suggests how to evaluate discriminant subspaces of LDA/GSVD. Actually, Howland and Park in [9] applied the corollary implicitly, but, in this paper, we explicitly connect the generalized singular value $\frac{\alpha}{\beta}$ with the generalized eigenvalue λ, the measure of discriminability.

Based on our analysis, the eigenvalue ratio $\frac{\lambda_b}{\lambda_w}$ and the square of the generalized singular value $\frac{\alpha^2}{\beta^2}$ are both equal to the generalized eigenvalue λ, the measure of discriminability. According to Fig. 1, Subspace 1, with the infinite eigenvalue ratio $\frac{\lambda_b}{\lambda_w}$, is the most discriminant subspace, followed by Subspace 2 and Subspace 3. However, Subspace 4 contains no discriminant information and can be safely thrown away. Therefore, the eigenvalue ratio $\frac{\lambda_b}{\lambda_w}$ or the generalized singular value $\frac{\alpha^2}{\beta^2}$ suggests how to choose the most discriminant subspaces.

3.4 Algorithm for LDA/FKT

Although we proved that FKT is equivalent to GSVD on the LDA problem, as we will see in Table 1, LDA/GSVD is computationally expensive. Since Subspace 4 contains no discriminant information, we may compute the Subspaces 1, 2, and 3 of LDA/FKT based on QR decomposition. Moreover, we use smaller matrices $\mathbf{H}_b$ and $\mathbf{H}_t$ because matrices $\mathbf{S}_b$, $\mathbf{S}_w$, and $\mathbf{S}_t$ may be too large to be formed. Our LDA/FKT algorithm is shown in Fig. 2.

Now, we analyze the computational complexity of the algorithm as follows:

1. **Time complexity.** Line 2 takes $O(DN^2)$ time to compute the QR decomposition on $\mathbf{H}_t$. To multiply two matrices, Line 3 takes $O(r_t^2 N)$ time, Line 4 takes $O(r_t DC)$ time, and Line 5 takes $O(r_t^2 C)$ time. Line 6 takes $O(r_t^3)$ time to invert $\widetilde{\mathbf{S}}_t$, multiply the matrices,

Input: The data matrix $\mathbf{A}$.
Output: Projection matrix Φ_F such that the J_F is maximized.
1) Compute $\mathbf{H}_b$ and $\mathbf{H}_t$ from data matrix $\mathbf{A}$ as in Equations (7) and (9).
2) Apply QR decomposition on $\mathbf{H}_t = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{D \times r_t}$, $\mathbf{R} \in \mathbb{R}^{r_t \times N}$ and $r_t = rank(\mathbf{H}_t)$.
3) Let $\widetilde{\mathbf{S}}_t = \mathbf{RR}^\top$, since $\widetilde{\mathbf{S}}_t = \mathbf{Q}^\top \mathbf{S}_t \mathbf{Q} = \mathbf{Q}^\top \mathbf{H}_t \mathbf{H}_t^\top \mathbf{Q} = \mathbf{RR}^\top$.
4) Let $\mathbf{Z} = \mathbf{Q}^\top \mathbf{H}_b$.
5) Let $\widetilde{\mathbf{S}}_b = \mathbf{ZZ}^\top$, since $\widetilde{\mathbf{S}}_b = \mathbf{Q}^\top \mathbf{S}_b \mathbf{Q} = \mathbf{Q}^\top \mathbf{H}_b \mathbf{H}_b^\top \mathbf{Q} = \mathbf{ZZ}^\top$.
6) Compute the eigenvectors $\{\mathbf{v}_i\}$ and eigenvalues $\{\lambda_i\}$ of $\widetilde{\mathbf{S}}_t^{-1} \widetilde{\mathbf{S}}_b$.
7) Sort the eigenvectors $\mathbf{v}_i$ according to λ_i in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > \lambda_{k+1} = \cdots = \lambda_{r_t} = 0$.
8) The final projection matrix $\Phi_F = \mathbf{QV}$, where $\mathbf{V} = \{\mathbf{v}_i\}$. Note that $\mathbf{QV}$ is the union of Subspaces 1, 2, and 3. If only Subspaces 1 and 2 are needed, $\Phi_F = \mathbf{QV}_k$ (the first k columns of $\mathbf{V}$).

Fig. 2. Algorithm 1: Apply QR decomposition to compute LDA/FKT.

and perform eigendecomposition on the $r_t \times r_t$ matrix $\widetilde{\mathbf{S}}_t^{-1}\widetilde{\mathbf{S}}_b$. Since $r_t < N$, $C \ll D$, the most intensive step is Line 2, which takes $O(DN^2)$ time to compute the QR decomposition. Thus, the time complexity is $O(DN^2)$.

2. **Space complexity.** Lines 2 and 4 involve matrices $\mathbf{H}_t$ and $\mathbf{H}_b$. Because of the size of the matrix, $\mathbf{H}_t$ requires $O(DN)$ space in memory, and $\mathbf{H}_b$ requires $O(DC)$. Lines 3, 5, and 6 only involve $\mathbf{R} \in \mathbb{R}^{r_t \times N}$, $\mathbf{Z} \in \mathbb{R}^{r_t \times C}$, and $\widetilde{\mathbf{S}}_t$, $\widetilde{\mathbf{S}}_b \in \mathbb{R}^{r_t \times r_t}$, which are all small matrices. Therefore, the space complexity is $O(DN)$.

4 Comparison

Although Fisherface, PCA+NULL, LDA/GSVD, and LDA/QR were all proposed independently and appear to be different algorithms, in this section, we explain how FKT provides insights into these methods.

4.1 Fisherface: Subspaces 2 and 3

In the Fisherface method, PCA is performed first so as to make $\mathbf{S}_w$ nonsingular. This is done by throwing away Subspaces 1 and 4 (Fig. 1). As a result, Subspace 1, the most discriminant subspace in terms of the Fisher Criterion, is discarded. Therefore, Fisherface operates only in Subspaces 2 and 3 and is suboptimal.

4.2 PCA + NULL Space: Subspace 1

Considering the discriminant information contained in the null space of $\mathbf{S}_w$, PCA+NULL Space first removes the null space of $\mathbf{S}_t$, which is Subspace 4 (Fig. 1). Now, only Subspaces 1, 2, and 3 are left. Second, within Subspace 1, the principal components of $\mathbf{S}_b$ are computed. Thus, only Subspace 1, the most discriminant feature space, is used.

Other null space methods have also been reported in the literature, such as Direct LDA [21] and NULL Space [3]. The criterion used in these methods is a modified version of the Fisher Criterion, namely,

$$\Phi_{opt} = \arg\max_{\Phi} \|\Phi^\top \mathbf{S}_b \Phi\| \qquad s.t. \|\Phi^\top \mathbf{S}_w \Phi\| = 0. \tag{25}$$

Equation (25) shows that Φ_{opt} is the set of eigenvectors associated with the zero eigenvalues of $\mathbf{S}_w$ and the maximum eigenvalues of $\mathbf{S}_b$. Based on the eigenvalue ratio from Fig. 1, this is Subspace 1. Thus, the PCA+NULL, Direct LDA, and NULL space methods all operate only in Subspace 1. However, as we will show in our experiments in Section 6.1, using Subspace 1 alone is sometimes not sufficient for good discrimination because Subspaces 2 and 3 may be necessary. In the worst case, when Subspace 1 does not exist,[1] these null space methods will fail.

4.3 LDA/QR: Subspaces 1 and 2

To circumvent the nonsingularity requirement of $\mathbf{S}_w$ and reduce the computation, a two-stage strategy is used in LDA/QR [20]. The eigenspace (corresponding to nonzero eigenvalues) of $\mathbf{S}_b$ is computed by applying QR on $\mathbf{H}_b$. In fact, this is Subspace 1 $\bigcup$ Subspace 2 (Fig. 1) because the eigenvalues of $\mathbf{S}_b$ associated with Subspaces 3 and 4 are all zero, which are thrown away by the QR decomposition. Then, the eigenvectors corresponding to the smallest eigenvalues of $\widetilde{\mathbf{S}}_b^{-1}\widetilde{\mathbf{S}}_w$ are computed, equivalently, computing the eigenvectors corresponding to the largest $\frac{\lambda_b}{\lambda_w}$. Note that $\widetilde{\mathbf{S}}_b^{-1}\widetilde{\mathbf{S}}_w$, rather than $\widetilde{\mathbf{S}}_w^{-1}\widetilde{\mathbf{S}}_b$, is eigendecomposed because $\widetilde{\mathbf{S}}_w^{-1}$ may still be singular (λ_w is zero within Subspace 1). Therefore, as Fig. 1 illustrated, Subspaces 1 and 2 are preserved by LDA/QR. This means that LDA/QR operates in Subspaces 1 and 2.

1. Subspace 1 will not exist if $\mathbf{S}_w$ is full rank and invertible. This can happen if there are enough training samples.

4.4 LDA/GSVD: Subspaces 1, 2, 3, and 4

Both LDA/GSVD and LDA/FKT simultaneously diagonalize two matrices, but, so far, nobody has investigated the relationship between these two methods. In this paper, one of our contributions is the proof that LDA/GSVD and LDA/FKT are equivalent (see Appendix A for the proof). More specifically, from the perspective of FKT, the $\mathbf{Y}$ and $\mathbf{Z}$ in LDA/GSVD (see (10) and (11)) are just arbitrary rotation matrices. The discriminant subspace of LDA/GSVD $\mathbf{X}$ is equal to $[\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp]$, where $\mathbf{U}_\perp$ is Subspace 4, and $\mathbf{U}$ is the union of Subspaces 1, 2, and 3. This means $\mathbf{X}$ contains Subspaces 1, 2, 3, and 4 (see Fig. 1). Therefore, the subspaces obtained by LDA/GSVD are exactly those obtained by LDA/FKT. However, LDA/GSVD is computationally expensive (Table 1). In Fig. 2, we presented an efficient algorithm to compute LDA/FKT. This is achieved by using QR decomposition on $\mathbf{S}_t$ to obtain Subspaces 1, 2, and 3. We do not have to compute for Subspace 4, since it contains no discriminant information.

5 Multiple Discriminant Analysis (MDA)

5.1 MDA/FKT

From the perspective of the Bayes Classifier, LDA is optimal only for two Gaussian distributions with equal covariance matrices [5], [7], and the Fisher Criterion has been extended to handle multiple Gaussian distributions or classes with unequal covariance matrices. This suggests that, for multiple Gaussian distributions or classes with unequal covariance matrices, LDA-based methods are not optimal with respect to the Bayes Classifier. The worst case occurs when all classes have the same mean. In this case, $\mathbf{S}_b = 0$ and all LDA-based methods will fail. Subspaces 1 and 2 do not exist and we are left with only Subspaces 3 and 4, which are less discriminative. To handle these problems, we cast the multiclass problem into a binary pattern classification problem by introducing $\Delta = \mathbf{a}_i - \mathbf{a}_j$ and defining the intraclass space $\Omega_I = \{(\mathbf{a}_i - \mathbf{a}_j) \mid L(\mathbf{a}_i) = L(\mathbf{a}_j)\}$, as well as the extraclass space $\Omega_E = \{(\mathbf{a}_i - \mathbf{a}_j) \mid L(\mathbf{a}_i) \neq L(\mathbf{a}_j)\}$, where $L(\mathbf{a}_i)$ is the class label of $\mathbf{a}_i$. This idea has been used by other researchers, for example, Moghaddam in [15]. The statistics of Ω_I and Ω_E are defined as follows:

$$\mathbf{m}_I = \mathbf{m}_E = 0, \tag{26}$$

$$\mathbf{\Sigma}_I = \mathbf{H}_I \mathbf{H}_I^\top = \frac{1}{N_I} \sum_{L(\mathbf{a}_i)=L(\mathbf{a}_j)} (\mathbf{a}_i - \mathbf{a}_j)(\mathbf{a}_i - \mathbf{a}_j)^\top, \tag{27}$$

$$\mathbf{\Sigma}_E = \mathbf{H}_E \mathbf{H}_E^\top = \frac{1}{N_E} \sum_{L(\mathbf{a}_i)\neq L(\mathbf{a}_j)} (\mathbf{a}_i - \mathbf{a}_j)(\mathbf{a}_i - \mathbf{a}_j)^\top. \tag{28}$$

Here, $N_I = \frac{1}{2}\sum n_i(n_i - 1)$ is the number of samples in Ω_I and $N_E = \sum_{L_i \neq L_j} n_i n_j$ is the number of samples in Ω_E. For example, if every class has the same number of training samples, $n_i = n$ for $i = 1, \ldots, C$, then $N_I = \frac{1}{2}N(n-1)$ and $N_E = \frac{1}{2}N(N-n)$. Note that, usually, $rank(\mathbf{\Sigma}_E)$ and $rank(\mathbf{\Sigma}_I)$

are both greater than $C-1$, where C is the number of classes. $\mathbf{H}_I$ and $\mathbf{H}_E$ are the precursor matrices of Σ_I and Σ_E given by

$$\mathbf{H}_I = \frac{1}{\sqrt{N_I}}[\ldots,(\mathbf{a}_i-\mathbf{a}_j),\ldots], \quad \forall i>j \quad \text{such that} \quad L(\mathbf{a}_i)=L(\mathbf{a}_j), \tag{29}$$

$$\mathbf{H}_E = \frac{1}{\sqrt{N_E}}[\ldots,(\mathbf{a}_i-\mathbf{a}_j),\ldots], \quad \forall i>j \quad \text{such that} \quad L(\mathbf{a}_i)\neq L(\mathbf{a}_j). \tag{30}$$

Our goal is now to find a subspace Φ in which Ω_I and Ω_E are as separable as possible. Here, the Bhattacharyya distance [5], [7] is used because it measures the overlap of any two probability density functions (pdf). If the pdfs are Gaussian, the Bhattacharyya distance is given analytically by

$$D_{bh} = \frac{1}{8}(\mathbf{m}_E-\mathbf{m}_I)^\top\left(\frac{\Sigma_E+\Sigma_I}{2}\right)^{-1}(\mathbf{m}_E-\mathbf{m}_I) + \frac{1}{2}\ln\frac{|\frac{\Sigma_E+\Sigma_I}{2}|}{\sqrt{|\Sigma_E|}\sqrt{|\Sigma_I|}}.$$

Since $\mathbf{m}_E=\mathbf{m}_I$, this simplifies to

$$D_{bh} = \frac{1}{2}\ln\frac{|\frac{\Sigma_E+\Sigma_I}{2}|}{\sqrt{|\Sigma_E|}\sqrt{|\Sigma_I|}} = \frac{1}{4}\left\{\ln|\Sigma_E^{-1}\Sigma_I+\Sigma_I^{-1}\Sigma_E+2\mathbf{I}|-D\ln 4\right\}. \tag{31}$$

The optimal subspace Φ can be computed by maximizing the new criterion [7], which is the Bhattacharyya distance:

$$J_{MDA} = \ln|(\Phi^\top\Sigma_E\Phi)^{-1}(\Phi^\top\Sigma_I\Phi)+ (\Phi^\top\Sigma_I\Phi)^{-1}(\Phi^\top\Sigma_E\Phi)+2\mathbf{I}_d|. \tag{32}$$

Here, $\mathbf{I}_d$ is the identity matrix in the low-dimensional space. Note that the Bhattacharyya distance may still be used even if the underlying distributions are not Gaussian [5]. It has been proven in [7] that, to maximize J_{MDA}, we must choose the generalized eigenvectors of the matrix pair (Σ_I,Σ_E) corresponding to the largest $\lambda+\frac{1}{\lambda}$, where λ is the generalized eigenvalue. When Σ_E or Σ_I is singular, we cannot obtain the subspace directly by matrix inversion because of the singularity problem. However, we can apply FKT to analyze Σ_E and Σ_I so that we can obtain four subspaces with two eigenvalue curves as in Fig. 1. Suppose λ_I and λ_E are the eigenvalues associated with the same eigenvector and λ is the generalized eigenvalue of matrix pair (Σ_I,Σ_E), then $\lambda=\frac{\lambda_I}{\lambda_E}$ and $\lambda_I+\lambda_E=1$. Now, to maximize J_{MDA} is to compute the eigenvectors corresponding the maximal $\lambda+\frac{1}{\lambda}$ or, equivalently, $\frac{\lambda_E}{\lambda_I}+\frac{\lambda_I}{\lambda_E}$. This is realized when $\lambda_I\to 0$ and $\lambda_E\to 1$ or $\lambda_I\to 1$ and $\lambda_E\to 0$. In Fig. 1, Subspace 1 $\bigcup$ Subspace 3 is the most discriminant subspace that satisfies the above criterion. The leftmost and rightmost part of Subspace 2 can provide additional discriminant information. The eigenvector of Subspace 2 corresponding to equal eigenvalues ($\lambda_I=\lambda_E=0.5$) is the least discriminative. To distinguish from LDA/FKT, we will call this technique MDA/FKT. The key difference between them is that MDA/FKT is optimal in terms of the Bhattacharyya distance, whereas LDA/FKT is optimal in terms of the Fisher Criterion.

Compared with other related work on MDA, our MDA/FKT has some unique features. First, the discriminant subspace obtained by MDA/FKT is optimal in terms of Bhattacharyya distance. Moghaddam [15] computed the top eigenvectors of Σ_E and Σ_I individually as the projection subspaces, which may not be discriminant. Second, our method finds the globally optimal subspace by analytically maximizing the Bhattacharyya distance, which is the error bound of the Bayes Classifier. Another method recently proposed by De la Torre Frade and Kanade [6] maximizes the Kullback-Leibler divergence, which does not relate to the Bayes Classifier. Their method finds a local optimum by using an iterative method. Finally, MDA/FKT can provide more than $C-1$ discriminant eigenvectors because usually the rank of Σ_E and Σ_I is greater than $C-1$, the upper bound of $rank(\mathbf{S}_b)$. By comparison, LDA-based methods can only provide $C-1$ discriminative eigenvectors because $C-1$ is the upper bound of $rank(\mathbf{S}_b)$.

5.2 Algorithm for MDA/FKT

Based on the new criterion (32), our analyses on FKT show that Subspaces 1 and 3 are the most discriminant subspaces (Fig. 1). However, we cannot directly work on $\Sigma_I\in\mathbb{R}^{D\times D}$ and $\Sigma_E\in\mathbb{R}^{D\times D}$, which may be singular or too large to be formed. An alternative is to use the precursor matrices $\mathbf{H}_I\in\mathbb{R}^{D\times N_I}$ and $\mathbf{H}_E\in\mathbb{R}^{D\times N_E}$. However, it is not efficient to use $\mathbf{H}_E$ as well because N_E is too large. As shown above, $N_I\propto N$, but $N_E\propto N^2$, where N is the total number of samples. Although $N\ll D$, N^2 could be close to D or even greater. For example, in our face recognition experiments, when $C=67$, $D=5,600$, and $n=2$ (two training samples per class), then $N=Cn=134$, $N_I=67$, and $N_E=8,844$. Σ_E is $5,600\times 5,600$ and $\mathbf{H}_E$ is $5,600\times 8,844$ in size.

Can we find an efficient way to obtain the Subspaces 1 and 3 of MDA/FKT without $\mathbf{H}_E$ or Σ_E? Yes. Based on the relationship between $\mathbf{S}_t$, Σ_I, and Σ_E, we devise a method that works with $\mathbf{H}_I\in\mathbb{R}^{D\times N_I}$ and $\mathbf{H}_t\in\mathbb{R}^{D\times N}$ only. Let us start with a lemma (see Appendix B for the proof.):

Lemma 2. *If $\mathbf{S}_t$ is the total scatter matrix defined in (5) and Σ_I and Σ_E are the covariance matrices of the intraclass and extraclass defined in (27) and (28), then $2N\mathbf{S}_t=N_I\Sigma_I+N_E\Sigma_E$, where N is the total number of samples, N_I is the number of intraclass samples, and N_E is the number of extraclass samples.*

Let us define $\Sigma'_I=\frac{N_I}{2N}\Sigma_I$ and $\Sigma'_E=\frac{N_E}{2N}\Sigma_E$, then $\mathbf{S}_t=\Sigma'_I+\Sigma'_E$. To efficiently compute the generalized eigenvalues and eigenvectors of (Σ_I,Σ_E), we need the following theorem:

Theorem 2. *If $(\lambda,\mathbf{v})$ is the dominant generalized eigenvalue and eigenvector of the matrix pair (Σ_I,Σ_E) and $(\lambda',\mathbf{v}')$ is the dominant generalized eigenvalue and eigenvector of matrix pair (Σ'_I,Σ'_E), then $\mathbf{v}=\mathbf{v}'$ and $\lambda=\frac{N_I}{N_E}\lambda'$.*

Proof. The generalized eigenvalue equation of matrix pair (Σ_I,Σ_E) is $\Sigma_E\mathbf{v}=\lambda\Sigma_I\mathbf{v}$. Since $\Sigma_I=\frac{2N}{N_I}\Sigma'_I$ and $\Sigma_E=\frac{2N}{N_E}\Sigma'_E$, we have

$$\frac{2N}{N_E}\Sigma'_E\mathbf{v} = \lambda\frac{2N}{N_I}\Sigma'_i\mathbf{v} \tag{33}$$

and, thus,

$$\Sigma'_E\mathbf{v} = \frac{N_E}{N_I}\lambda\Sigma'_i\mathbf{v}. \tag{34}$$

Input: The data matrix $\mathbf{A}$.
Output: Projection matrix Φ_{MDA} such that the J_{MDA} is maximized.

1) Compute $\mathbf{H}_I$ and $\mathbf{H}_t$ from data matrix $\mathbf{A}$ as in Equations (30) and (9).
2) Apply QR decomposition on $\mathbf{H}_t = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{D\times r_t}, \mathbf{R} \in \mathbb{R}^{r_t\times N}$ and $r_t = rank(\mathbf{H}_t)$.
3) Let $\widetilde{\mathbf{S}}_t = \mathbf{R}\mathbf{R}^\top$, since $\widetilde{\mathbf{S}}_t = \mathbf{Q}^\top\mathbf{S}_t\mathbf{Q} = \mathbf{Q}^\top\mathbf{H}_t\mathbf{H}_t^\top\mathbf{Q} = \mathbf{R}\mathbf{R}^\top$.
4) Let $\mathbf{Z} = \mathbf{Q}^\top\mathbf{H}_I$.
5) Let $\widetilde{\Sigma}'_I = \frac{N_I}{2N}\mathbf{Z}\mathbf{Z}^\top$, since $\widetilde{\Sigma}'_I = \mathbf{Q}^\top\Sigma'_I\mathbf{Q} = \frac{N_I}{2N}\mathbf{Q}^\top\mathbf{H}_I\mathbf{H}_I^\top\mathbf{Q} = \frac{N_I}{2N}\mathbf{Z}\mathbf{Z}^\top$.
6) Compute the eigenvectors $\{\mathbf{v}_i\}$ and eigenvalues $\{\sigma_i\}$ of $\widetilde{\mathbf{S}}_t^{-1}\widetilde{\Sigma}'_I$.
7) Compute the generalized eigenvalues $\{\lambda_i\}$ of (Σ_I, Σ_E) using $\lambda_i = \frac{N_I\sigma_i}{N_E(1-\sigma_i)}$.
8) Sort the eigenvectors $\mathbf{v}_i$ according to $\lambda_i + \frac{1}{\lambda_i}$ in decreasing order.
9) The final projection matrix $\Phi_{MDA} = \mathbf{QV}_k$ (the first k columns of $\mathbf{V}$), where $\mathbf{V} = \{\mathbf{v}_i\}$. Note that k could be greater than $C-1$.

Fig. 3. Algorithm 2: Apply QR decomposition to compute MDA/FKT.

Comparing with the generalized eigenvalue equation of matrix pair (Σ'_E, Σ'_I): $\Sigma'_E\mathbf{v}' = \lambda'\Sigma'_I\mathbf{v}'$, we observe that $\mathbf{v} = \mathbf{v}'$ and $\lambda = \frac{N_I}{N_E}\lambda'$. □

Furthermore, it is obvious that the proof of Theorem 2 is valid for the other generalized eigenvalues and eigenvectors as well, not just the dominant one. There is a one-to-one mapping between the corresponding eigenvalues and eigenvectors of the two pairs of matrices.

Therefore, to compute any generalized eigenvalue and eigenvector of the matrix pair (Σ_I, Σ_E), we can work on the matrix pair (Σ'_I, Σ'_E). Since $\mathbf{S}_t = \Sigma'_I + \Sigma'_E$ and, based on our FKT analysis, each generalized eigenvalue λ' is equal to $\frac{\lambda'_I}{\lambda'_E}$, the eigenvalue ratio of FKT. This can be realized by using smaller matrices $\mathbf{H}_t \in \mathbb{R}^{D\times N}$ and $\mathbf{H}_I \in \mathbb{R}^{D\times N_I}$, where $N \ll D$ and $N_I \ll D$.

Now, the idea of our algorithm is similar to LDA/FKT, which applies QR decomposition on $\mathbf{H}_b$ and $\mathbf{H}_t$ to obtain Subspaces 1, 2, and 3. However, we apply QR decomposition on $\mathbf{H}_t$ and $\mathbf{H}_I$ to compute Subspaces 1, 2, and 3. Note that MDA/FKT is optimal for the Bhattacharyya distance, whereas LDA/FKT is optimal for the Fisher Criterion. Our MDA/FKT algorithm is shown in Fig. 3.

5.3 Computational Complexity

Let us analyze the time complexity first. Line 2 takes $O(DN^2)$ time to compute the QR decomposition on $\mathbf{H}_t$. To multiply two matrices, Line 3 takes $O(r_t^2N)$ time, Line 4 takes $O(r_tDN_I)$ time, and Line 5 takes $O(r_t^2N_I)$ time. Line 6 takes $O(r_t^3)$ time to invert $\widetilde{\mathbf{S}}_t$, multiply the matrices, and perform eigendecomposition on the $r_t \times r_t$ matrix $\widetilde{\mathbf{S}}_t^{-1}\widetilde{\Sigma}_I$. Since $r_t < N$, $N_I \ll D$, the most intensive step is Line 2, which takes $O(DN^2)$ time to compute the QR decomposition. Thus, the time complexity is $O(DN^2)$.

Considering the space complexity, in MDA/FKT, Lines 2 and 4 involve matrices $\mathbf{H}_t$ and $\mathbf{H}_I$. Each matrix requires $O(DN)$ space in the memory because of the size of the matrix. Lines 3, 5, and 6 only involve $\mathbf{R} \in \mathbb{R}^{r_t\times N}, \mathbf{Z} \in \mathbb{R}^{r_t\times N_I}$ and $\widetilde{\mathbf{S}}_t, \widetilde{\Sigma}'_I \in \mathbb{R}^{r_t\times r_t}$, which are all small matrices. Therefore, the space complexity is $O(DN)$.

Table 1 compares the time/space complexity of the methods mentioned in this paper. Observe that our MDA/FKT is comparable to most of LDA-based methods. MDA/FKT, however, is optimal in terms of Bhattacharyya distance, the error bound of the Bayes Classifier, which is not the case for other methods.

TABLE 2
Statistics of Our Data Sets

Type	Dataset	Size	Dimension	# of classes
Synthetic	Toy 1	300	3	3
	Toy 2	125	3	2
Real data	MFEAT	2000	649	10
	PIE	1608	5600	67
	Banca	6240	2805	52

6 Experiments

Up until now, we have shown that FKT can be used to unify other LDA-based methods. Moreover, we proposed a new approach for MDA. In this section, we evaluate the performance of LDA/FKT and MDA/FKT by using synthetic and real data. The synthetic data has two sets, whereas the real data consists of three sets for digit recognition and face recognition. Table 2 shows the statistics of the data sets in our experiments.

The experimental setting for recognition is described as follows. For PCA, we take the top $C-1$ principal components,[2] where C is the number of classes. For Fisherface, we apply PCA first and take 100 principal components, followed by LDA. For MDA/FKT, when performing comparison with other methods, we project MDA/FKT to $C-1$-dimensional space. With respect to recognition, we employ 1-NN in the low-dimensional space for all methods in these experiments.

6.1 Toy Problems

To evaluate the performance of MDA/FKT, we begin with two toy examples:

- **Toy 1:** *Three Gaussian classes: same mean, different covariance matrices*. The three classes share the same zero mean in 3D space and each class has 100 points. They have different covariance matrices:

$$\mathbf{C}_1 = [1,1,0]^\top \times [1,1,0] + 0.1[0,1,1]^\top \times [0,1,1],$$

$$\mathbf{C}_2 = [0,1,1]^\top \times [0,1,1] + 0.1[1,0,1]^\top \times [1,0,1],$$

and

$$\mathbf{C}_3 = [1,0,1]^\top \times [1,0,1] + 0.1[1,1,0]^\top \times [1,1,0]$$

(Fig. 4a). Note that LDA-based methods will fail here because $\mathbf{S}_b = 0$.
- **Toy 2:** *Two classes: Gaussian mixture*. We also have two classes in 3D space. One class contains 50 points and the other contains 75. The first class is generated from a single Gaussian with zero mean and $0.5\mathbf{I}$ covariance. The second class is a Gaussian mixture which consists of three components with different means: $[1,4,0]$, $[2\sqrt{3},-2,0]$, and $[-2\sqrt{3},-2,0]$. Each component has

2. Asymptotically, if PCA can extract more features than LDA, it will perform better.

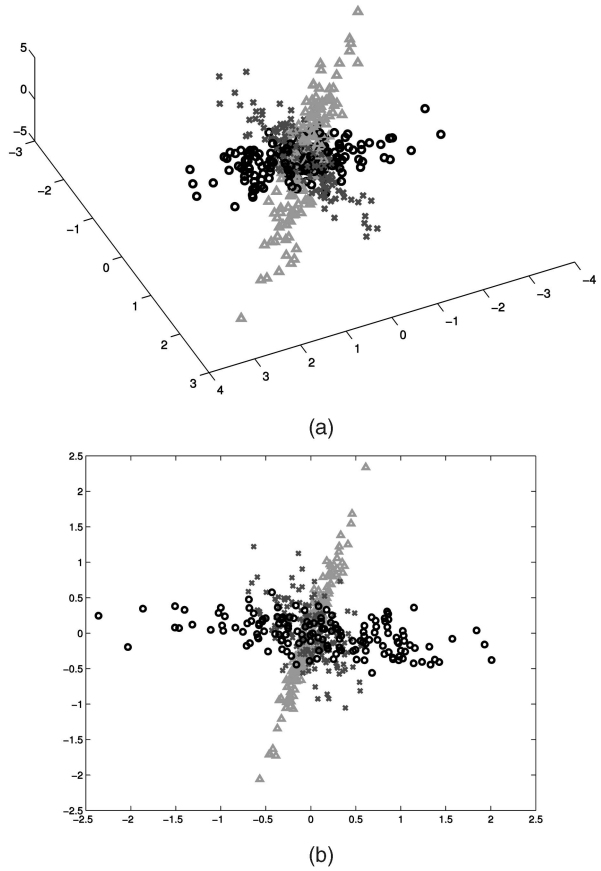


Fig. 4. (a) Original 3D data. (b) Two-dimensional projection by MDA/FKT. Different colors (or markers) represent different classes. Note that the original 3D data share the same mean ($\mathbf{S}_b = 0$). This results in the failure of LDA-based methods.

25 points and $0.5\mathbf{I}$ covariance, and its mixture proportion is $\frac{1}{3}$, as shown in Fig. 5a.

In Toy 1, it does not make sense to maximize trace$(\mathbf{S}_w^{-1}\mathbf{S}_b)$ because $\mathbf{S}_b = 0$. This means that LDA fails when the different classes share the same mean. However, MDA/FKT can still be applied because $\Sigma_I \neq 0$ and $\Sigma_E \neq 0$. As shown in Fig. 4b, by using MDA/FKT, we can obtain a 2D discriminative subspace, which can still approximate the structure of the original 3D data. Even though it is hard to determine the decision boundaries in Fig. 4b, we can discern that the three classes lie principally along three different axes. In Toy 2, LDA/FKT can obtain only a one-dimensional (1D) projection (Fig. 5c) because, for two classes, $rank(\mathbf{S}_b) = 1$. Note that the 1D projections of LDA/FKT overlap significantly, which means it is hard to do classification. However, as shown in Fig. 5b, MDA/FKT can obtain a 2D subspace because the rank of Σ_I or Σ_E depends on the number of samples, not the number of classes. The larger discriminative subspace of MDA/FKT makes it possible to separate the classes.

Let us summarize these two toy problems: First, MDA/FKT can still work even if all classes share the same mean, whereas all LDA-based methods fail because $\mathbf{S}_b = 0$. Second, MDA/FKT can provide larger discriminative subspaces than LDA-based methods because the latter ones are limited by the number of classes.

6.2 Digit Recognition

We perform digit recognition to compare MDA/FKT with LDA-based methods on the MFEAT [12] data set. This consists of handwritten digits ("0"-"9") (10 classes) with 649-dimensional features. These features comprise six feature sets: Fourier coefficients, profile correlations, Karhunen-Love coefficients, pixel averages in 2×3 windows, Zernike moments, and morphological features. For each class, we have 200 patterns; 30 of them are chosen randomly as training samples and the rest for testing. To evaluate the stability of each method, we repeat the sampling 10 times so that we can compute the mean and standard deviation of the recognition accuracy.

As shown in Fig. 6a, the accuracy of LDA/FKT and MDA/FKT is about 95 percent with small standard deviations, which means an accurate and stable performance. For MDA/FKT, we can investigate the relationship between performance and the projected dimension. Fig. 6b shows a plot of accuracy versus projected dimensions. The accuracy reaches the maximum when the projected dimension is around 8, after which it remains flat even if we increase the projected dimension. This suggests that MDA/FKT reaches its best performance around eight dimensions. Another observation is that the projected dimension should not be limited by the number of classes. For example, here, we have $C = 10$ classes, but Fig. 6b illustrates that seven or eight dimensions can give almost the same accuracy as $C - 1 = 9$ projected dimensions.

6.3 Face Recognition

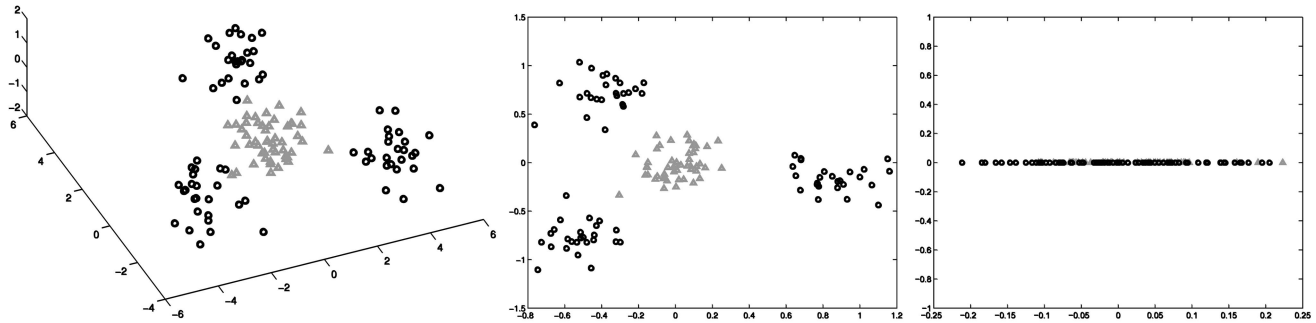
We also perform experiments on real data on two face data sets:

1. **PIE face data set** [17]. We choose 67 subjects[3] and each subject has 24 frontal face images taken under room lighting. All of these face images are aligned based on eye coordinates and cropped to 70×80. Fig. 7a shows a sample of PIE face images used in our experiments. The major challenge in this data set is to do face recognition under different illuminations.
2. **Banca face data set** [1]. This contains 52 subjects and each subject has 120 face images, which are normalized to 51×55 in size. By using a Webcam and an expensive camera, these subjects were recorded in three different scenarios over a period of three months. Each face image contains illumination, expression, and pose variations because the subjects are required to talk during the recording (Fig. 7b).

Fig. 7 shows a sample of PIE and Banca face images used in our experiments.

For face recognition, usually, we have an undersampled problem, which is also the reason for the singularity of $\mathbf{S}_w$. To evaluate the performance under such a situation, we randomly choose N training samples from each subject, $N = 2, \cdots, 12$, and the remaining images are used for testing. For each set of N training samples, we employ cross validation so that we can compute the mean and standard deviation for classification accuracies. We show the mean and standard deviation (in parenthesis) of the recognition rate from 10 runs (see Table 3). Note that the largest number in each row is highlighted in bold.

3. The PIE data set contains 68 subjects altogether. We omitted one because s/he has too few frontal face images.

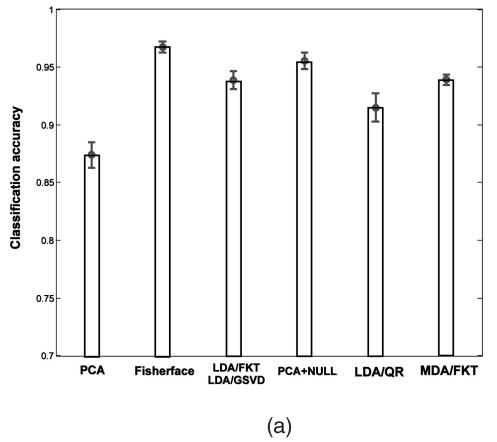


(a) (b) (c)

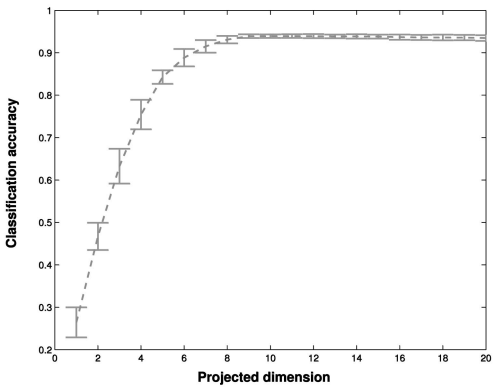
Fig. 5. (a) Original 3D data. (b) Two-dimensional projection by MDA/FKT. (c) One-dimensional projection by LDA/FKT. The projection of MDA/FKT is more separable than that of LDA/FKT because the former can provide a larger discriminant subspace.

As shown in Table 3, we observe that the more training samples, the better the recognition accuracy. To be more specific, on both data sets, for each method, increasing the number of training samples increases the mean recognition rate and decreases the standard deviation. Note that, when the training set is small, LDA/FKT significantly outperforms the other methods. For example, for the PIE data set, with two training samples, LDA/FKT achieves about 98 percent accuracy compared with the next highest of 90 percent from PCA+NULL (see the first row of Table 3). Moreover, with four training samples, LDA/FKT achieves about 100 percent compared with 94 percent of the PCA+NULL method (see the second row of Table 3). The standard deviation of LDA/FKT is also significantly smaller than that of the other methods. For the Banca data set, we have a similar observation. With two and four training samples, LDA/FKT achieves about 5 percent higher in accuracy than the next highest (see the seventh and eighth rows of Table 3). This shows that LDA/FKT can handle small-sample-size problems very well. With more training samples, that is, 6-12, LDA/FKT is not the best but falls behind the highest by no more than 1.8 percent (see the bottom four rows of Table 3). One possible reason is that LDA/FKT is optimal as a linear classifier, whereas, for the Banca data set, the face images under expression and pose variations are nonlinearly distributed.

To compare PCA, MDA/FKT, and various LDA-based methods with different training samples, we also visualize the average classification accuracies in Fig. 8.

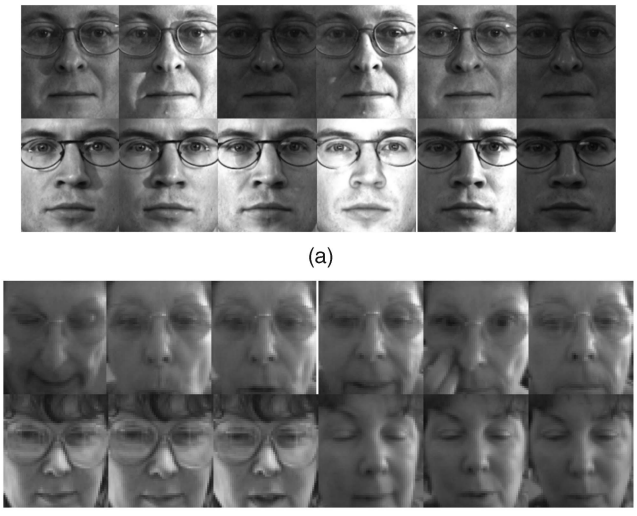


(a)



(b)

Fig. 6. The accuracy rate (mean and standard deviation) of digit recognition by using: (a) PCA, MDA, and LDA-based methods. (b) MDA/FKT: accuracy versus projected dimension.



(a) (b)

Fig. 7. A sample of face images: (a) PIE data set and (b) Banca data set. Each row represents one subject. Note that the PIE face images have only illumination variation, whereas the Banca ones have illumination, pose, and expression variations. Moreover, the Banca images are captured in different scenarios with different cameras.

TABLE 3
Classification Accuracy (%) of Different Methods with Different Training Set Sizes

Dataset	# of training samples	PCA	Fisherface	PCA+NULL	LDA/QR	LDA/FKT LDA/GSVD	MDA/FKT
PIE	2	39.62(1.56)	88.95(1.23)	89.49(1.15)	85.74(1.66)	**98.09**(0.71)	95.03(1.16)
	4	59.60(2.07)	93.58(0.85)	93.78(0.94)	91.05(2.48)	**99.74**(0.32)	99.06(0.57)
	6	73.53(1.53)	96.18(0.38)	96.05(0.53)	94.70(1.60)	**99.96**(0.04)	99.77(0.23)
	8	82.61(1.62)	97.78(0.59)	97.41(0.63)	95.58(1.23)	**99.97**(0.03)	99.95(0.05)
	10	88.10(1.46)	98.58(0.66)	98.27(0.70)	97.57(1.39)	**100.0**(0.00)	99.91(0.08)
	12	92.76(1.46)	98.82(0.48)	98.52(0.50)	98.34(0.61)	**99.98**(0.02)	99.95(0.04)
Banca	2	38.30(1.57)	65.33(3.71)	65.62(2.72)	55.04(1.92)	**70.45**(1.74)	70.42(1.74)
	4	54.45(1.78)	79.81(2.19)	68.32(2.09)	73.59(4.84)	85.33(1.67)	**86.15**(1.31)
	6	65.37(0.74)	90.15(0.94)	**91.41**(0.85)	80.23(4.73)	90.87(0.77)	91.06(1.13)
	8	73.63(1.04)	94.20(0.85)	**94.33**(0.56)	87.40(5.15)	93.85(0.56)	93.55(0.62)
	10	78.54(1.33)	**95.93**(0.79)	94.97(0.56)	88.52(3.10)	94.35(0.60)	94.77(0.56)
	12	82.34(1.31)	**96.92**(0.48)	96.01(0.44)	91.30(3.41)	95.12(0.35)	95.35(0.27)

Note that, since MDA/FKT is not limited by the number of classes (unlike LDA), we may project the data onto a space whose dimension is greater than the number of classes. Fig. 9 shows a plot of accuracy versus projected dimensions for different numbers of training samples for both the PIE and Banca data sets. We observe the following:

1. The more training samples we use, the better the recognition rate. This is consistent with our experiments (in Table 3) and it has also been confirmed by other researchers. The reason is that a larger set of

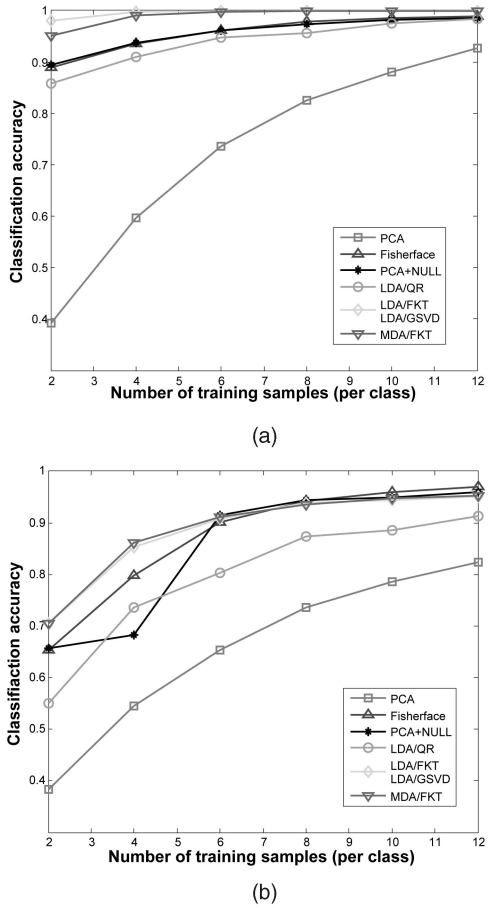


Fig. 8. Face recognition by using PCA, MDA/FKT, and various LDA-based methods with different numbers of training samples per class. (a) PIE (67 classes). (b) Banca (52 classes). Note that, for visualization, we only show the mean of the accuracies from 10 runs. These plots are generated from Table 3.

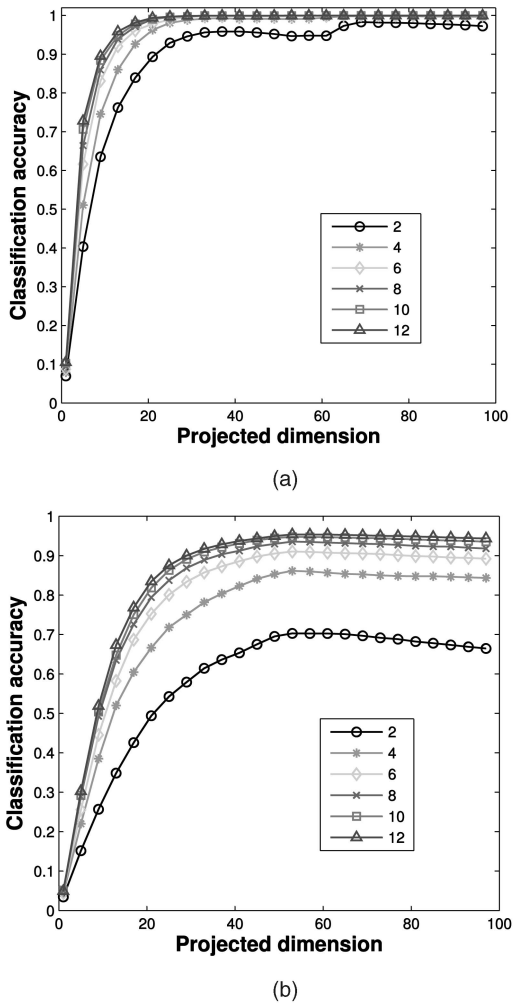


Fig. 9. Face recognition by varying the number of training samples per class and the projected dimension. (a) MDA/FKT curves on PIE (67 classes). (b) MDA/FKT curves on Banca (52 classes). Note that, for visualization, we only show the mean of the accuracies from 10 runs.

training data can sample the underlying distribution more accurately than a smaller set.
2. With fewer training samples per class, say, two or four, the highest accuracy is obtained at $C-1$ projected dimensions ($C = 67$ for PIE, $C = 52$ for Banca).
3. However, with more training samples per class, say, six or more, we can obtain a high classification rate with fewer than $C-1$ projected dimensions. For example, on the PIE data set, with eight or 10 training samples per class, we can obtain 98 percent accuracy by using only 30 projected dimensions (Fig. 9a). The curves remain flat with increasing dimensions. Thus, there is no incentive to use more than 30 dimensions.

Now, let us summarize our experiments on real data sets. First, MDA/FKT is comparable to LDA-based methods with respect to the accuracy. Second, LDA/FKT and MDA/FKT significantly outperform other LDA-based methods for small-sample-size problems.

7 Conclusion

In this paper, we showed how FKT can provide valuable insights into LDA. We derived and proved the relationship between GSVD, FKT, and LDA and then unified different LDA-based methods. Furthermore, we proposed a new method—MDA/FKT—to handle the MDA problem. More precisely:

1. We decomposed the whole data space into four subspaces by using FKT. For the LDA problem, we proved that the GSVD is equivalent to the FKT.
2. We proved that the eigenvalue ratio of FKT and the square of the generalized singular value of GSVD are equal to the generalized eigenvalue of LDA, which is the measure of discriminability according to the Fisher Criterion. It unifies these three methods that were previously separately proposed.
3. We proposed a unified framework to understanding different methods, that is, Fisherface, PCA+NULL, LDA/QR, and LDA/GSVD. Our theoretical analyses showed how to choose the discriminant subspaces based on the generalized eigenvalue, the essential measure of separability.
4. We also compared some common LDA methods with LDA/FKT. Most of these methods are suboptimal in terms of the Fisher Criterion. More specifically, Fisherface, PCA+NULL, and LDA/QR all operate in different parts of the discriminative subspaces of LDA/FKT. We showed that LDA/GSVD and LDA/FKT are, in fact, equivalent, but our LDA/FKT is more efficient than LDA/GSVD with respect to computation.
5. We further presented MDA/FKT with the following properties:
   a. It is derived from the Bhattacharyya distance, which is the error bound of the Bayes Classifier. This is theoretically superior to the Fisher Criterion, which is based on scatter matrices and which does not relate to the Bayes Classifier.
   b. It can provide larger discriminative subspaces; in contrast, LDA-based methods are limited by the number of classes.
   c. It works even if $\mathbf{S}_b = 0$, which is where LDA-based methods fail. Furthermore, for Gaussian mixture pdf, it works better than LDA.
   d. It can be realized by an efficient algorithm. This algorithm is comparable to most of LDA-based methods with respect to computation and storage.
6. We experimentally showed the superiority of LDA/FKT and MDA/FKT. In particular, for small-sample-size problems, LDA/FKT and MDA/FKT work significantly better than other methods. In the case of MDA/FKT, we further observed that using a small projected subspace (dimension $\approx \frac{C-1}{2}$) is enough to achieve high accuracy when the training set is sufficiently large.

An interesting future work is to extend our theory to a nonlinear discriminant analysis. One way is to use the kernel trick employed in support vector machines (SVMs), for example, construct kernelized between-class scatter and within-class scatter matrices. FKT may yet again reveal new insights into the kernelized LDA.

Appendix A
Proof of Lemma 1

Proof. GSVD$\Longrightarrow$FKT.

Based on GSVD,

$$\mathbf{S}_b = \mathbf{H}_b\mathbf{H}_b^\top = \mathbf{X}^{-\top}\begin{bmatrix}\Sigma_b^\top\Sigma_b & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{X}^{-1}, \tag{35}$$

$$\mathbf{S}_w = \mathbf{H}_w\mathbf{H}_w^\top = \mathbf{X}^{-\top}\begin{bmatrix}\Sigma_w^\top\Sigma_w & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{X}^{-1}. \tag{36}$$

Thus,

$$\mathbf{X}^\top(\mathbf{S}_b+\mathbf{S}_w)\mathbf{X} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}. \tag{37}$$

Since $\Sigma_b^\top\Sigma_b + \Sigma_w^\top\Sigma_w = \mathbf{I} \in \mathbb{R}^{r_t\times r_t}$, if we choose the first r_t columns of $\mathbf{X}$ as $\mathbf{P}$, that is, $\mathbf{P} = \mathbf{X}_{(d,r_t)}$, then $\mathbf{P}^\top(\mathbf{S}_b+\mathbf{S}_w)\mathbf{P} = \mathbf{I}$. This is exactly FKT. Meanwhile, we can obtain that $\Lambda_b = \Sigma_b^\top\Sigma_b$ and $\Lambda_w = \Sigma_w^\top\Sigma_w$.

FKT$\Longrightarrow$GSVD.

Based on FKT $\mathbf{P} = \mathbf{U}\mathbf{D}^{-1/2}$,

$$\widetilde{\mathbf{S}}_b = \mathbf{P}^\top\mathbf{S}_b\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{U}^\top\mathbf{H}_b\mathbf{H}_b^\top\mathbf{U}\mathbf{D}^{-1/2}, \tag{38}$$

$$\widetilde{\mathbf{S}}_b = \mathbf{V}\Lambda_b\mathbf{V}^\top. \tag{39}$$

Hence,

$$\mathbf{D}^{-1/2}\mathbf{U}^\top\mathbf{H}_b\mathbf{H}_b^\top\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{V}\Lambda_b\mathbf{V}^\top. \tag{40}$$

In general, there is no unique decomposition on the above equation because $\mathbf{H}_b\mathbf{H}_b^\top = \mathbf{H}_b\mathbf{Y}\mathbf{Y}^\top\mathbf{H}_b^\top$ for any orthogonal matrix $\mathbf{Y} \in \mathbb{R}^{C\times C}$. That is,

$$\mathbf{D}^{-1/2}\mathbf{U}^\top\mathbf{H}_b\mathbf{Y}\mathbf{Y}^\top\mathbf{H}_b^\top\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{V}\Lambda_b\mathbf{V}^\top, \tag{41}$$

$$\mathbf{Y}^\top\mathbf{H}_b^\top\mathbf{U}\mathbf{D}^{-1/2} = \widehat{\Sigma}_b\mathbf{V}^\top, \tag{42}$$

$$\mathbf{Y}^\top\mathbf{H}_b^\top\mathbf{U}\mathbf{D}^{-1/2}V = \widehat{\Sigma}_b, \tag{43}$$

where $\widehat{\Sigma}_b \in \mathbb{R}^{C\times r_t}$ and $\Lambda_b = \widehat{\Sigma}_b^\top\widehat{\Sigma}_b$. If we define $\mathbf{X} = [\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp] \in \mathbb{R}^{d\times d}$. Then,

$$\begin{aligned}\mathbf{Y}^\top\mathbf{H}_b^\top\mathbf{X} &= \mathbf{Y}^\top\mathbf{H}_b^\top[\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp], &(44)\\ &= [\mathbf{Y}^\top\mathbf{H}_b^\top\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{0}], &(45)\\ &= [\widehat{\Sigma}_b, \mathbf{0}]. &(46)\end{aligned}$$

Here, $\mathbf{H}_b^\top\mathbf{U}_\perp = \mathbf{0}$, and $\mathbf{H}_w^\top\mathbf{U}_\perp = \mathbf{0}$ because $\mathbf{U}_\perp$ is the intersection of the null space of $\mathbf{S}_b$ and $\mathbf{S}_w$. Similarly, we can get $\mathbf{Z}^\top\mathbf{H}_w^\top\mathbf{X} = [\widehat{\Sigma}_w, \mathbf{0}]$, where $\mathbf{Z} \in \mathbb{R}^{N\times N}$ is an arbitrary orthogonal matrix, $\widehat{\Sigma}_w \in \mathbb{R}^{N\times r_t}$, and $\Lambda_w = \widehat{\Sigma}_w^\top\widehat{\Sigma}_w$. Since $\Lambda_b + \Lambda_w = \mathbf{I}$ and $\mathbf{I} \in \mathbb{R}^{r_t\times r_t}$ is an identity matrix, it is easy to check that $\widehat{\Sigma}_b^\top\widehat{\Sigma}_b + \widehat{\Sigma}_w^\top\widehat{\Sigma}_w = \mathbf{I}$, which satisfies the constraint of GSVD.

Now, we have to prove $\mathbf{X}$ is nonsingular

$$\begin{aligned}\mathbf{X}\mathbf{X}^\top &= [\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}, \mathbf{U}_\perp]\begin{bmatrix}\mathbf{V}^\top\mathbf{D}^{-1/2}\mathbf{U}\top\\ \mathbf{U}_\perp^\top\end{bmatrix}\\ &= \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{U}_\perp^\top &(47)\\ &= [\mathbf{U}, \mathbf{U}_\perp]\begin{bmatrix}\mathbf{D}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}^\top\\ \mathbf{U}_\perp^\top\end{bmatrix}.\end{aligned}$$

Here, $\mathbf{V} \in \mathbb{R}^{r\times r}$ and $[\mathbf{U}, \mathbf{U}_\perp]$ are orthogonal matrices. Note that $\mathbf{U}^\top\mathbf{U}_\perp = \mathbf{0}$ and $\mathbf{U}_\perp^\top\mathbf{U} = \mathbf{0}$. From the above equation, $\mathbf{X}\mathbf{X}^\top$ can be eigendecomposed with positive eigenvalues, which means $\mathbf{X}$ is also nonsingular. This completes the proof. □

APPENDIX B
PROOF OF LEMMA 2

Proof. Since

$$\begin{aligned}\mathbf{\Sigma}_i &= \frac{1}{N_I}\sum_{L(\mathbf{a}_I)=L(\mathbf{a}_j)}(\mathbf{a}_i - \mathbf{a}_j)(\mathbf{a}_i - \mathbf{a}_j)^\top,\\ \mathbf{\Sigma}_E &= \frac{1}{N_E}\sum_{L(\mathbf{a}_i)\neq L(\mathbf{a}_j)}(\mathbf{a}_i - \mathbf{a}_j)(\mathbf{a}_i - \mathbf{a}_j)^\top.\end{aligned}$$

Then,

$$\begin{aligned}N_I\Sigma_I + N_E\Sigma_E &= \sum_i\sum_j(\mathbf{a}_i - \mathbf{a}_j)(\mathbf{a}_i - \mathbf{a}_j)^\top\\ &= \sum_i\sum_j\left(\mathbf{a}_i\mathbf{a}_i^\top - \mathbf{a}_i\mathbf{a}_j^\top - \mathbf{a}_j\mathbf{a}_i^\top + \mathbf{a}_j\mathbf{a}_j^\top\right)\\ &= 2N\sum_i\left(\mathbf{a}_i\mathbf{a}_i^\top\right) - 2\Big(\sum\mathbf{a}_i\Big)\Big(\sum\mathbf{a}_j^\top\Big)\\ &= 2N\sum_i\left(\mathbf{a}_i\mathbf{a}_i^\top\right) - 2N^2\mathbf{m}\mathbf{m}^\top\\ &= 2N\Big(\sum\mathbf{a}_i\mathbf{a}_i^\top - N\mathbf{m}\mathbf{m}^\top\Big),\end{aligned} \quad (48)$$

where $\mathbf{m} = \frac{1}{N}\sum\mathbf{a}_i$ is the total mean of the samples.

On the other hand,

$$\begin{aligned}\mathbf{S}_t &= \sum_i^N(\mathbf{a}_i - \mathbf{m})(\mathbf{a}_i - \mathbf{m})^\top\\ &= \sum\left(\mathbf{a}_i\mathbf{a}_i^\top - \mathbf{a}_i\mathbf{m}^\top - \mathbf{m}\mathbf{a}_i^\top + \mathbf{m}\mathbf{m}^\top\right)\\ &= \sum\mathbf{a}_i\mathbf{a}_i^\top - N\mathbf{m}\mathbf{m}^\top.\end{aligned} \quad (49)$$

By examining (48) and (49), we can see that $2N\mathbf{S}_t = N_I\Sigma_I + N_E\Sigma_E$. □

REFERENCES

[1] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruíz, and J.P. Thiran, "The Banca Database and Evaluation Protocol," *Proc. Fourth Int'l Conf. Audio and Video-Based Biometric Person Authentication*, pp. 625-638, 2003.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces versus Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, July 1997.

[3] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminant Common Vectors for Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4-13, Jan. 2005.

[4] S. Chakrabarti, S. Roy, and M. Soundalgekar, "Fast and Accurate Text Classification via Multiple Linear Discriminant Projections," *Proc. Int'l Conf. Very Large Data Bases*, pp. 658-669, 2002.

[5] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, second ed. John Wiley & Sons, 2000.

[6] F. de la Torre Frade and T. Kanade, "Multimodal Oriented Discriminant Analysis," *Proc. Int'l Conf. Machine Learning*, Aug. 2005.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. Academic Press, 1990.

[8] G.H. Golub and C.F. Van Loan, *Matrix Computations*, third ed. Johns Hopkins Univ. Press, 1996.

[9] P. Howland and H. Park, "Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 995-1006, Aug. 2004.

[10] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the Small Sample Size Problem of LDA," *Proc. IEEE Int'l Conf. Pattern Recognition*, vol. 3, pp. 29-32, 2002.

[11] X. Huo, "A Statistical Analysis of Fukunaga-Koontz Transform," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 123-126, Feb. 2004.

[12] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[13] I.T. Jolliffe, *Principal Component Analysis*. Springer, 1986.

[14] C.F. Van Loan, "Generalizing the Singular Value Decomposition," *SIAM J. Numerical Analysis*, vol. 13, no. 1, pp. 76-83, 1976.

[15] B. Moghaddam, "Principal Manifolds and Probabilistic Subspaces for Visual Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 780-788, June 2002.

[16] C.C. Paige and M.A. Saunders, "Towards a Generalized Singular Value Decomposition," *SIAM J. Numerical Analysis*, vol. 18, no. 3, pp. 398-405, 1981.

[17] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615-1618, Dec. 2003.

[18] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[19] X. Wang and X. Tang, "Dual-Space Linear Discriminant Analysis for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 564-569, June 2004.

[20] J. Ye and Q. Li, "A Two-Stage Linear Discriminant Analysis via QR-Decomposition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 929-942, June 2005.

[21] H. Yu and H. Yang, "A Direct LDA Algorithm for High-Dimensional Data—with Application to Face Recognition," *Pattern Recognition*, vol. 34, no. 10, pp. 2067-2070, 2001.

[22] S. Zhang and T. Sim, "When Fisher Meets Fukunaga-Koontz: A New Look at Linear Discriminants," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 323-329, June 2006.

Sheng Zhang received the bachelor's degree from Zhejiang University, China, in 1998, the Master's degree from the Institute of Automation, Chinese Academy of Sciences, in 2001, and the PhD degree from the School of Computing, National University of Singapore in 2006. He now works in the Department of Electrical and Computer Engineering, University of California at Santa Barbara (UCSB) as a postdoctoral researcher. His research interests include face recognition, computer vision, statistical pattern recognition, and machine learning. He is a member of the IEEE and the IEEE Computer Society.

Terence Sim received the BS degree in computer science and engineering from the Massachusetts Institute of Technology in 1990, the MS degree in computer science from Stanford University in 1991, and the PhD degree in electrical and computer engineering from Carnegie Mellon University in 2002. He is an assistant professor in the School of Computing, National University of Singapore. His research interests are in biometrics, face recognition, computer vision, digital photography, and music processing. He chairs the Workgroup on Cross-Jurisdictional and Societal Aspects in the Biometrics Technical Committee, Singapore. He is a member of the IEEE.