Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2156–2172
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based?

Wei-Shi Zheng[a,c], J.H. Lai[b,c,*], Stan Z. Li[d]

[a]School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou, PR China
[b]Department of Electronics and Communication Engineering, School of Information Science and Technology, Sun Yat-sen University, Guangzhou, PR China
[c]Guangdong Province Key Laboratory of Information Security, PR China
[d]Center for Biometrics and Security Research and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, PR China

Received 18 August 2006; received in revised form 27 November 2007; accepted 29 November 2007

Abstract

Recent advances have shown that algorithms with (2D) matrix-based representation perform better than the traditional (1D) vector-based ones. In particular, 2D-LDA has been widely reported to outperform 1D-LDA. However, would the matrix-based linear discriminant analysis be always superior and when would 1D-LDA be better? In this paper, we investigate into these questions and have a comprehensive comparison between 1D-LDA and 2D-LDA in theory and in experiments. We analyze the heteroscedastic problem in 2D-LDA and formulate mathematical equalities to explore the relationship between 1D-LDA and 2D-LDA; then we point out potential problems in 2D-LDA. It is shown that 2D-LDA has eliminated the information contained in the covariance information between different local geometric structures, such as the rows or the columns, which is useful for discriminant feature extraction, whereas 1D-LDA could preserve such information. Interestingly, this new finding indicates that 1D-LDA is able to gain higher Fisher score than 2D-LDA in some extreme case. Furthermore, sufficient conditions on which 2D-LDA would be Bayes optimal for two-class classification problem are derived and comparison with 1D-LDA in this aspect is also analyzed. This could help understand how 2D-LDA is expected to achieve at its best, further discover its relationship with 1D-LDA, and well support other findings. After the theoretical analysis, comprehensive experimental results are reported by fairly and extensively comparing 1D-LDA with 2D-LDA. In contrast to the existing view that some 2D-LDA based algorithms would perform better than 1D-LDA when the number of training samples for each class is small or when the number of discriminant features used is small, we show that it is not always true and show that some standard 1D-LDA based algorithms could perform better in those cases on some challenging data sets.
© 2007 Elsevier Ltd. All rights reserved.

Keywords: Fisher's linear discriminant analysis (LDA); Matrix-based representation; Vector-based representation; Pattern recognition

1. Introduction

Over the last two decades, many subspace algorithms have been developed for feature extraction. Among them are principal component analysis (PCA) [1–4], (Fisher's) linear discriminant analysis (LDA) [4–8], independent component analysis (ICA) [9–12], non-negative matrix factorization (NMF) [13–15], locality preserving projection [16] and Bayesian probabilistic subspace [17,18], etc.

Most well-known subspace methods require the input patterns to be shaped in vector form. Recently there are efforts seeking to extract features directly without any vectorization work on image samples, i.e., the representation of an image sample is retained in matrix form. Based on this idea, some well-known algorithms are developed, including two-dimensional principal component analysis (2D-PCA) [19,20] and two-dimensional linear discriminant analysis (2D-LDA) [21–23].

2D-PCA was first proposed by Yang et al. [19,20], and a generalized work has been subsequently described in [24]

* Corresponding author. Department of Electronics and Communication Engineering, School of Information Science and Technology, Sun Yat-sen University, Guangzhou, PR China. Tel.: +86 020 84035440.
E-mail addresses: wszheng@ieee.org (W.-S. Zheng), stsljh@mail.sysu.edu.cn (J.H. Lai), szli@nlpr.ia.ac.cn (S.Z. Li).

0031-3203/$30.00 © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2007.11.025

called bilateral-projection-based 2DPCA (B2DPCA). Ye then proposed the generalized low rank approximations of matrices (GLRAM) [25] as a further development of 2D-PCA. Recently a modification on 2D-PCA was proposed in Ref. [26] and it could be treated as implementing 2D-PCA after rearrangement of the entries of an image matrix.

For supervised learning, 2D-LDA has also been developed recently. Xiong et al. [22] and Li et al. [21] extended one-dimensional LDA (1D-LDA), a vector-based scheme, to 2D-LDA. In contrast to [21,22] which only do transform on one side of the image matrix, i.e., either left side or right side, some methods have been proposed for extraction of the discriminative transforms on both sides of the image matrix. Yang et al. [27] proposed to do the IMLDA (uncorrelated image matrix-based LDA) twice, i.e., IMLDA is first implemented to find the optimal discriminant projection on the right side of the matrix and then to find another optimal discriminant projection on the left side. Similarly, Kong et al. [28] proposed to first extract the 2D-LDA discriminative projections on both sides of the image matrix independently and then combine them by some processing. Different from them, Ye et al. proposed an iterative scheme to extract the transforms on both sides [23] simultaneously. Recently, some other modifications on 2D-LDA [29–31] are proposed. Especially, in Ref. [30], similar to Fisherface [8], 2D-LDA is processed after the implementation of 2D-PCA. Though such rapid development appeared in the last two years; however, Liu et al. [32] actually had suggested a 2D image matrix-based (Fisher's) linear discriminant technique which performed LDA directly on image matrices in 1993. In nature, the idea behind is to construct the covariance matrix, including total-class scatter matrix, within-class scatter matrix and between-class scatter matrix, by just using the original image samples represented in matrix form. Moreover, some recent studies [24,28,33,34] have realized that two-dimensional matrix-based algorithms are special blocked-based methods such as column-based or row-based LDA\PCA in essence.

2D-LDA is attractive since it is efficient in computation and always avoids the "small sample size problem" [8,35–38] that the within-class scatter matrix is always singular in 1D-LDA when the training sample size is (much) smaller than the dimensionality of the data. Recently, the 2D-LDA based algorithms have been experimentally reported superior to some standard 1D-LDA based algorithms, such as Fisherface [8], on some limited data sets.

However, one may ask: "Could 2D-LDA always perform the best?" "Why would it be better sometimes?" "Is there any drawback in 2D-LDA?" "What is the intrinsic relationship between 1D-LDA and 2D-LDA?" "1D-LDA is Bayes optimal for two-class classification under some sufficient conditions, and then what is the situation for 2D-LDA? What are the differences between 1D-LDA and 2D-LDA under their sufficient conditions being Bayes optimal?" After all, "When is 1D-LDA better than 2D-LDA?"

We do investigation into these questions and present an extensive analysis between 1D-LDA and 2D-LDA in theory and in experiments. This is, to the best of our knowledge, the first of such attempt with comprehensive study. The contributions of this paper are summarized as follows:

(1) Extensive theoretical comparisons between 1D-LDA and 2D-LDA are presented, and we have the following findings:
   (a) From the statistical point of view, 2D-LDA would also be confronted with the "Heteroscedastic Problem" and the problem would be more serious for 2D-LDA than the one for 1D-LDA.
   (b) Mathematical equalities are formulated to explore the relationship between 1D-LDA and 2D-LDA. It gives a novel way to show that 2D-LDA loses the covariance information among different local geometry structures in the image such as rows or columns, while 1D-LDA could preserve those relations for feature extraction. It then breaks the appearance view that 2D-LDA is able to utilize the global geometry structure of an image. Interestingly, we further find that 1D-LDA is able to achieve higher Fisher score than 2D-LDA in some extreme case as shown in the paper.
   (c) The sufficient conditions when 2D-LDA is Bayes optimal for two-class classification problem are given and proved. They could help give an interpretation what 2D-LDA is expected ideally. Moreover further discussions between 1D-LDA and 2D-LDA are presented when those sufficient conditions are satisfied or not.
(2) Extensive experiments are conducted to compare 1D-LDA with 2D-LDA. The experimental results break the existing views and indeed show that 2D-LDA would not always be superior to 1D-LDA when the number of training samples for each class is small or when the number of discriminant features used is small.

Though this paper focuses on (Fisher's) LDA; however, the analysis could be useful for other similar algorithms. The remainder of this paper is outlined as follows. In Section 2, a brief review of 1D-LDA and 2D-LDA is given. In Section 3, theoretical analysis between 1D-LDA and 2D-LDA is presented. In Section 4, extensive experiments are conducted. Finally, we have a summarization in Section 5.

2. Reviews

2.1. Notations

Suppose $\{(\mathbf{x}_1^1, \mathbf{X}_1^1, C_1), \ldots, (\mathbf{x}_{N_1}^1, \mathbf{X}_{N_1}^1, C_1), \ldots, (\mathbf{x}_1^L, \mathbf{X}_1^L, C_L), \ldots, (\mathbf{x}_{N_L}^L, \mathbf{X}_{N_L}^L, C_L)\}$ are image samples from L classes. The n-dimensional vector $\mathbf{x}_i^k \in \mathbf{R}^n$ is the ith sample of the kth class C_k and $\mathbf{X}_i^k \in \mathbf{R}^{row \times col}$ is its corresponding $row \times col$ image matrix, where $i = 1, \ldots, N_k$ and N_k is the number of training samples of class C_k. Let $N = \sum_{j=1}^{L} N_j$ be the total sample size. Define $\mathbf{u}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{x}_i^k$ as the mean vector of samples of class C_k and $\mathbf{U}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{X}_i^k$ as its corresponding mean matrix. Let $\mathbf{u} = \sum_{k=1}^{L} \frac{N_k}{N} \mathbf{u}_k$ be the mean vector of all samples and $\mathbf{U} = \sum_{k=1}^{L} \frac{N_k}{N} \mathbf{U}_k$ be its corresponding mean matrix.

2.2. 1D-LDA (one-dimensional LDA)

1D-LDA aims to find the discriminative vector $\mathbf{w}_{opt}$ such that

$$\mathbf{w}_{opt} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w}}, \tag{1}$$

where $\mathbf{S}_b=\sum_{k=1}^{L}\frac{N_k}{N}(\mathbf{u}_k-\mathbf{u})(\mathbf{u}_k-\mathbf{u})^{\mathrm{T}}$, $\mathbf{S}_w=\frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}(\mathbf{x}_i^k-\mathbf{u}_k)(\mathbf{x}_i^k-\mathbf{u}_k)^{\mathrm{T}}=\sum_{k=1}^{L}\frac{N_k}{N}\mathbf{S}_w^k$, $\mathbf{S}_w^k=\frac{1}{N_k}\sum_{i=1}^{N_k}(\mathbf{x}_i^k-\mathbf{u}_k)(\mathbf{x}_i^k-\mathbf{u}_k)^{\mathrm{T}}$ are between-class scatter matrix, within-class scatter matrix and within-class scatter matrix of class C_k, respectively. In practice, due to the curse of high dimensionality, $\mathbf{S}_w$ is always singular. So far, some well-known standard variations of 1D-LDA have been developed to overcome this problem, such as Fisherface [8] and its further developments [40,41], Nullspace LDA [35–37], direct LDA [42], LDA/QR [38,43] and regularized LDA [5,44–47], etc. Thereof, regularized LDA is always implemented as follows:

$$\mathbf{w}_{r\text{-}opt} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}}{\mathbf{w}^{\mathrm{T}}(\mathbf{S}_w+\lambda\mathbf{I})\mathbf{w}}, \quad \lambda>0. \tag{2}$$

Other efforts are also made for obtaining more discriminative and robust 1D-LDA algorithms in the small sample size case, such as constraint-based LDA algorithm [48,49], weight-based LDA algorithm [50], mixture model-based LDA [51], locally LDA [52] and oriented LDA [53], etc.

2.3. 2D-LDA (two-dimensional LDA)

2D-LDA directly performs discriminant feature analysis on an image matrix rather than on a vector. 2D-LDA tries to find the optimal vector $\mathbf{w}_{opt}^{2d}$ such that

$$\mathbf{w}_{opt}^{2d} = \arg\max_{\mathbf{w}^{2d}} \frac{\mathbf{w}^{2d\mathrm{T}}\mathbf{S}_b^{2d}\mathbf{w}^{2d}}{\mathbf{w}^{2d\mathrm{T}}\mathbf{S}_w^{2d}\mathbf{w}^{2d}}, \tag{3}$$

where $\mathbf{S}_b^{2d}=\sum_{k=1}^{L}\frac{N_k}{N}(\mathbf{U}_k-\mathbf{U})(\mathbf{U}_k-\mathbf{U})^{\mathrm{T}}$ and $\mathbf{S}_w^{2d}=\frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}(\mathbf{X}_i^k-\mathbf{U}_k)(\mathbf{X}_i^k-\mathbf{U}_k)^{\mathrm{T}}$ are between-class scatter matrix and within-class scatter matrix, respectively. An alternative approach of 2D-LDA could be driven by the following criterion:

$$\tilde{\mathbf{w}}_{opt}^{2d} = \arg\max_{\tilde{\mathbf{w}}^{2d}} \frac{\tilde{\mathbf{w}}^{2d\mathrm{T}}\tilde{\mathbf{S}}_b^{2d}\tilde{\mathbf{w}}^{2d}}{\tilde{\mathbf{w}}^{2d\mathrm{T}}\tilde{\mathbf{S}}_w^{2d}\tilde{\mathbf{w}}^{2d}}, \tag{4}$$

where $\tilde{\mathbf{S}}_b^{2d}=\sum_{k=1}^{L}\frac{N_k}{N}(\mathbf{U}_k-\mathbf{U})^{\mathrm{T}}(\mathbf{U}_k-\mathbf{U})$ and $\tilde{\mathbf{S}}_w^{2d}=\frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}(\mathbf{X}_i^k-\mathbf{U}_k)^{\mathrm{T}}(\mathbf{X}_i^k-\mathbf{U}_k)$.

Equality (Criterion) (3) or (4) is called the unilateral 2D-LDA [28]. As aforementioned, a generalization of 2D-LDA called the bilateral 2D-LDA (B-2D-LDA) [23,28] finds a pair discriminant vectors $(\mathbf{w}_{l\text{-}opt}^{2d}, \mathbf{w}_{r\text{-}opt}^{2d})$ satisfying:

$$(\mathbf{w}_{l\text{-}opt}^{2d}, \mathbf{w}_{r\text{-}opt}^{2d}) = \arg\max_{(\mathbf{w}_l^{2d},\mathbf{w}_r^{2d})} \frac{\sum_{k=1}^{L}\frac{N_k}{N}\mathbf{w}_l^{2d\mathrm{T}}(\mathbf{U}_k-\mathbf{U})\mathbf{w}_r^{2d}\mathbf{w}_r^{2d\mathrm{T}}(\mathbf{U}_k-\mathbf{U})^{\mathrm{T}}\mathbf{w}_l^{2d}}{\frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}\mathbf{w}_l^{2d\mathrm{T}}(\mathbf{X}_i^k-\mathbf{U}_k)\mathbf{w}_r^{2d}\mathbf{w}_r^{2d\mathrm{T}}(\mathbf{X}_i^k-\mathbf{U}_k)^{\mathrm{T}}\mathbf{w}_l^{2d}}. \tag{5}$$

3. 1D-LDA vs. 2D-LDA: theoretical analysis

In this part, to compare with 1D-LDA, we first mainly focus on 2D-LDA in terms of equality (3). It does not mean the comparison would lose the generality. It is because equality (4) would become equality (3) if the input matrices are transposed first, and also so far it is hard to obtain a closed form solution but a practical solution [23,28,54] is popular and always found for equality (5). Analysis will be extended to the variations of 2D-LDA in terms of equalities (4)–(5) in Section 3.4.

Without loss of generality, define $\mathbf{X}_i^k=[\mathbf{X}_i^k(1),\mathbf{X}_i^k(2),\ldots,\mathbf{X}_i^k(col)]\in\mathbf{R}^{row\times col}$ and its corresponding vector form $\mathbf{x}_i^k=[\mathbf{X}_i^k(1)^{\mathrm{T}},\mathbf{X}_i^k(2)^{\mathrm{T}},\ldots,\mathbf{X}_i^k(col)^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{X}_i^k(j)\in\mathbf{R}^{row\times 1}$ is the jth column of matrix $\mathbf{X}_i^k$. We then have

$$\begin{aligned}\mathbf{U}_k &= [\mathbf{U}_k(1),\ldots,\mathbf{U}_k(col)]\\ &= \left[\frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{X}_i^k(1),\ldots,\frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{X}_i^k(col)\right],\\ \mathbf{U} &= [\mathbf{U}(1),\ldots,\mathbf{U}(col)]\\ &= \left[\sum_{k=1}^{L}\frac{N_k}{N}\mathbf{U}_k(1),\ldots,\sum_{k=1}^{L}\frac{N_k}{N}\mathbf{U}_k(col)\right],\\ \mathbf{u}_k &= [\mathbf{U}_k(1)^{\mathrm{T}},\ldots,\mathbf{U}_k(col)^{\mathrm{T}}]^{\mathrm{T}},\\ \mathbf{u} &= [\mathbf{U}(1)^{\mathrm{T}},\ldots,\mathbf{U}(col)^{\mathrm{T}}]^{\mathrm{T}}.\end{aligned}$$

As indicated in Refs. [28,33], it is easy to verify the following:

$$\begin{aligned}\mathbf{S}_b^{2d} &= \sum_{k=1}^{L}\frac{N_k}{N}\sum_{j=1}^{col}(\mathbf{U}_k(j)-\mathbf{U}(j))(\mathbf{U}_k(j)-\mathbf{U}(j))^{\mathrm{T}}\\ &= \mathbf{S}_{b,1}^{2d}+\cdots+\mathbf{S}_{b,col}^{2d},\end{aligned} \tag{6}$$

$$\begin{aligned}\mathbf{S}_w^{2d} &= \frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}\sum_{j=1}^{col}(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))^{\mathrm{T}}\\ &= \mathbf{S}_{w,1}^{2d}+\cdots+\mathbf{S}_{w,col}^{2d},\end{aligned} \tag{7}$$

where

$$\begin{aligned}\mathbf{S}_{b,j}^{2d} &= \sum_{k=1}^{L}\frac{N_k}{N}(\mathbf{U}_k(j)-\mathbf{U}(j))(\mathbf{U}_k(j)-\mathbf{U}(j))^{\mathrm{T}},\\ &\quad j=1,\ldots,col,\\ \mathbf{S}_{w,j}^{2d} &= \frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))^{\mathrm{T}},\\ &\quad j=1,\ldots,col.\end{aligned}$$

3.1. Heteroscedastic problem

First the 2D-LDA criterion in terms of equality (3) could be equivalently written as

$$\mathbf{w}_{opt}^{2d} = \arg\max_{\mathbf{w}^{2d}} \frac{\mathbf{w}^{2d\mathrm{T}}\left\{\frac{1}{col}\sum_{j=1}^{col}\mathbf{S}_{b,j}^{2d}\right\}\mathbf{w}^{2d}}{\mathbf{w}^{2d\mathrm{T}}\left\{\frac{1}{col}\sum_{j=1}^{col}\mathbf{S}_{w,j}^{2d}\right\}\mathbf{w}^{2d}}.$$

It can be found that the between-class information of 2D-LDA in terms of equality (3) is modeled by averaging all between-class scatter matrices $\mathbf{S}_{b,j}^{2d}$ with respect to different column indexes and models the within-class information similarly by averaging all $\mathbf{S}_{w,j}^{2d}$. From the statistical point of view, both $\mathbf{S}_b^{2d}$ and $\mathbf{S}_w^{2d}$ are "plug-in" estimates according to equalities (6)–(7). However, if columns with different indexes of images are heteroscedastic in essence, i.e., $\mathbf{S}_{b,j}^{2d} \neq \mathbf{S}_{b,i}^{2d}, \forall i \neq j$ or $\mathbf{S}_{w,j}^{2d} \neq \mathbf{S}_{w,i}^{2d}, \forall i \neq j$, then those "plug-in" estimates $\mathbf{S}_b^{2d}$ and $\mathbf{S}_w^{2d}$ would be inappropriate if the differences between $\mathbf{S}_{b,j}^{2d}$ or the differences between $\mathbf{S}_{w,j}^{2d}$ are significantly large. In such case the heteroscedastic problem [39] has to be addressed. We note that 1D-LDA would also be confronted with the heteroscedastic problem when the covariance matrices of different classes, i.e., $\mathbf{S}_w^k$, $k=1,\ldots,L$, are not equal [39], and it breaks the assumption of LDA that within-class covariance matrices of all classes are equal. However, the problem for 2D-LDA is different from the one for 1D-LDA in the following aspects. It is observed that samples learned by 2D-LDA in terms of equality (3) are actually the columns of images according to equalities (6)–(7), while columns are always obviously different if they are not coherent. Hence, on one hand, for estimation of within-class scatter information, columns with different indexes of images within the same class could be heteroscedastic (i.e., $\mathbf{S}_{w,j}^{2d}$ are not equal), even if the image samples in vector form are not heteroscedastic (i.e., $\mathbf{S}_w^k$ are equal). On the other hand, the heteroscedastic problem in 1D-LDA is mainly due to the unequal within-class covariance matrices of different classes, but such a problem could additionally happen to $\mathbf{S}_b^{2d}$ in 2D-LDA for estimation of between-class scatter information, because it is formulated by averaging all $\mathbf{S}_{b,j}^{2d}$. Therefore, it would be expected that the heteroscedastic problem in 2D-LDA could be more serious than that in 1D-LDA. However, such a seriously potential problem in 2D-LDA has not been pointed out before.

3.2. Relationship between 1D-LDA and 2D-LDA

Let $\mathbf{w}=[\widehat{\mathbf{w}}_1^{\mathrm{T}},\ldots,\widehat{\mathbf{w}}_{col}^{\mathrm{T}}]^{\mathrm{T}}$ be any n-dimensional vector, where $\widehat{\mathbf{w}}_i \in \mathbf{R}^{row\times 1}$. To explore the relationship between 1D-LDA and 2D-LDA, we first have the following lemma, and its proof can be found in Appendix A.

Lemma 1. *If* $\widehat{\mathbf{w}}_1,\ldots,\widehat{\mathbf{w}}_{col} \in \mathbf{R}^{row\times 1}$ *are imposed to be equivalent, i.e.,*

$$\mathbf{w}^{2d}=\widehat{\mathbf{w}}_1=\cdots=\widehat{\mathbf{w}}_{col}\in\mathbf{R}^{row\times 1}, \tag{8}$$

then the following relations are valid:

$$\tilde{\mathbf{w}}^{\mathrm{T}}\mathbf{S}_b\tilde{\mathbf{w}}=\mathbf{w}^{2d^{\mathrm{T}}}\mathbf{S}_b^{2d}\mathbf{w}^{2d}+\mathbf{w}^{2d^{\mathrm{T}}}\left\{\sum_{k=1}^{L}\frac{N_k}{N}\sum_{j=1,h=1,j\neq h}^{col}(\mathbf{U}_k(j)-\mathbf{U}(j))(\mathbf{U}_k(h)-\mathbf{U}(h))^{\mathrm{T}}\right\}\mathbf{w}^{2d}, \tag{9}$$

$$\tilde{\mathbf{w}}^{\mathrm{T}}\mathbf{S}_w\tilde{\mathbf{w}}=\mathbf{w}^{2d^{\mathrm{T}}}\mathbf{S}_w^{2d}\mathbf{w}^{2d}+\mathbf{w}^{2d^{\mathrm{T}}}\left\{\frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}\sum_{j=1,h=1,j\neq h}^{col}(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))(\mathbf{X}_i^k(h)-\mathbf{U}_k(h))^{\mathrm{T}}\right\}\mathbf{w}^{2d}, \tag{10}$$

where

$$\tilde{\mathbf{w}}=\left[\underbrace{\mathbf{w}^{2d^{\mathrm{T}}},\ldots,\mathbf{w}^{2d^{\mathrm{T}}}}_{col}\right]^{\mathrm{T}}. \tag{11}$$

2D-LDA is apparently indicated to preserve global geometric information of image since it directly lies on samples represented in image matrix form. However, the above lemma reveals that unlike 1D-LDA, it may lose the covariance information among different local geometry structures, such as the columns here. This is because in equalities (9) and (10), summation of the covariance information of data after a 2D-LDA transform and the eliminated covariance information by 2D-LDA between different local geometry structures is just the covariance information of data after a special 1D-LDA transform, where $\mathbf{w}^{2d^{\mathrm{T}}}\mathbf{S}_b^{2d}\mathbf{w}^{2d}$ is the between-class covariance information and $\mathbf{w}^{2d^{\mathrm{T}}}\mathbf{S}_w^{2d}\mathbf{w}^{2d}$ is the within-class covariance information induced by the 2D-LDA transform $\mathbf{w}^{2d}$. Hence 2D-LDA does not completely utilize global geometric information of an image. Though $\tilde{\mathbf{w}}$ is a special $row\cdot col(=n)$ dimensional vector; however, equalities (9)–(10) suggest 1D-LDA could preserve those information.

Although some recent studies [28,33] have indicated that 2D-LDA is a special block-based algorithm; however, the relationship between 1D-LDA and 2D-LDA has not been further explored theoretically as shown in equalities (9) and (10) before. Based on them, we here provide a new way to reveal that those part-based local geometric structures are considered separately and show the covariance information between them is not taken into account by 2D-LDA in theory.

Furthermore, the relationship formulated by Lemma 1 could in fact provide a more in-depth insight view. The following theorem then tells such an interesting issue.

Theorem 1. *1 D-LDA can have higher Fisher score than* 2*D-LDA if the following cases are valid*:

$$\sum_{k=1}^{L}\frac{N_k}{N}\sum_{j=1,h=1,j\neq h}^{col}(\mathbf{U}_k(j)-\mathbf{U}(j))(\mathbf{U}_k(h)-\mathbf{U}(h))^{\mathrm{T}}=\mathbf{0}, \tag{12}$$

$$\frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}\sum_{j=1,h=1,j\neq h}^{col}(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))(\mathbf{X}_i^k(h)-\mathbf{U}_k(h))^{\mathrm{T}}=\mathbf{0}. \tag{13}$$

Proof. In such a case, the following relations hold:

$$\tilde{\mathbf{w}}^{\mathrm{T}}\mathbf{S}_b\tilde{\mathbf{w}}=\mathbf{w}^{2d^{\mathrm{T}}}\mathbf{S}_b^{2d}\mathbf{w}^{2d}, \tag{14}$$

$$\tilde{\mathbf{w}}^{\mathrm{T}}\mathbf{S}_w\tilde{\mathbf{w}}=\mathbf{w}^{2d^{\mathrm{T}}}\mathbf{S}_w^{2d}\mathbf{w}^{2d}. \tag{15}$$

Since $\tilde{\mathbf{w}}$ is just a special n-dimensional vector, hence it is valid that:

$$\max_{\mathbf{w}^{2d}\in\mathbf{R}^{row}} \frac{\mathbf{w}^{2d\mathrm{T}}\mathbf{S}_b^{2d}\mathbf{w}^{2d}}{\mathbf{w}^{2d\mathrm{T}}\mathbf{S}_w^{2d}\mathbf{w}^{2d}} \leqslant \max_{\mathbf{w}\in\mathbf{R}^{n}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w}}. \tag{16}$$

That is, 1D-LDA can obtain higher Fisher score than 2D-LDA. □

One situation when equalities (12) and (13) are valid is the case that columns with different indexes of image matrices are statistically independent. A further interpretation of equality (16) in such case could be provided from another point of view in next section.

3.3. 2D-LDA: a Bayes optimal feature extractor under sufficient conditions

It is known that for two-class classification problem 1D-LDA will be Bayes optimal if data are normally distributed with equal covariance matrices within each class [4,5]. Then what is the situation for 2D-LDA? The analysis here attempts to seek the sufficient conditions when 2D-LDA would be Bayes optimal for two-class classification. Finally, the differences between 1D-LDA and 2D-LDA will be discussed when those sufficient conditions are satisfied or not.

Suppose $\mathbf{X}=[\mathbf{X}(1),\ldots,\mathbf{X}(col)]$ is a random $\mathbf{R}^{row\times col}$ matrix, where $\mathbf{X}(j)\in\mathbf{R}^{row}$, $j=1,\ldots,col$. Let $p(\mathbf{X})$ and $p(\mathbf{X}(j))$ be the probability density functions of $\mathbf{X}$ and $\mathbf{X}(j)$, respectively, and let $p(\mathbf{X}|C_k)$and $p(\mathbf{X}(j)|C_k)$ be the class-conditional probability density functions of class C_k. Then it is valid that

$$p(\mathbf{X}) = p(\mathbf{X}(1),\ldots,\mathbf{X}(col)),$$
$$p(\mathbf{X}|C_k) = p(\mathbf{X}(1),\ldots,\mathbf{X}(col)|C_k).$$

If $\mathbf{X}(1),\ldots,\mathbf{X}(col)$ are independent, we then have

$$p(\mathbf{X}) = \prod_{j=1}^{col} p(\mathbf{X}(j)), \quad p(\mathbf{X}|C_k) = \prod_{j=1}^{col} p(\mathbf{X}(j)|C_k). \tag{17}$$

Given two classes C_1 and C_2, to classify $\mathbf{X}$ using Bayesian decision principle, it is said $\mathbf{X}\in C_1$ if and only if $p(C_1|\mathbf{X}) > p(C_2|\mathbf{X})$ else $\mathbf{X}\in C_2$. Note that $P(C_k|\mathbf{X}) = \frac{p(\mathbf{X}|C_k)P(C_k)}{p(\mathbf{X})}$, where $P(C_k)$ is the prior probability of class C_k. If $\mathbf{X}(1),\ldots,\mathbf{X}(col)$ are assumed to be independent,[1] then

$$P(C_k|\mathbf{X}) = \prod_{j=1}^{col} \frac{p(\mathbf{X}(j)|C_k)}{p(\mathbf{X}(j))} P(C_k), \tag{18}$$

$$\log(P(C_k|\mathbf{X})) = \sum_{j=1}^{col}\{\log(p(\mathbf{X}(j)|C_k)) - \log(p(\mathbf{X}(j)))\} + \log(P(C_k)). \tag{19}$$

[1]This condition could be strict and a discussion will be given at the end of this section.

If all the jth columns $\mathbf{X}(j)$ of the kth class C_k are normally distributed with mean $\mathbf{M}_k(j)$ and covariance matrix $\mathbf{\Sigma}_k^j$, i.e.,

$$p(\mathbf{X}(j)|C_k) = (2\pi)^{-row/2}|\mathbf{\Sigma}_k^j|^{-1/2}\exp\{-\tfrac{1}{2}(\mathbf{X}(j) - \mathbf{M}_k(j))^{\mathrm{T}}(\mathbf{\Sigma}_k^j)^{-1}(\mathbf{X}(j)-\mathbf{M}_k(j))\},$$

$$\log(p(\mathbf{X}(j)|C_k)) = -\frac{row}{2}\log 2\pi - \tfrac{1}{2}\log|\mathbf{\Sigma}_k^j| - \tfrac{1}{2}(\mathbf{X}(j) - \mathbf{M}_k(j))^{\mathrm{T}}(\mathbf{\Sigma}_k^j)^{-1}(\mathbf{X}(j)-\mathbf{M}_k(j)) \tag{20}$$

then the Bayes classifier function $g_k(\mathbf{X})$ can be formulated as

$$\begin{aligned} g_k(\mathbf{X}) &= \log(P(C_k|\mathbf{X})) \\ &= \sum_{j=1}^{col}\left\{-\frac{row}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}_k^j| - \frac{1}{2}(\mathbf{X}(j) - \mathbf{M}_k(j))^{\mathrm{T}}(\mathbf{\Sigma}_k^j)^{-1}(\mathbf{X}(j)-\mathbf{M}_k(j)) - \log(p(\mathbf{X}(j)))\right\} \\ &\quad + \log(P(C_k)). \end{aligned} \tag{21}$$

In practice, utilizing the maximum likelihood principle, $\mathbf{M}_k(j)$ and $\mathbf{\Sigma}_k^j$ could be estimated by

$$\hat{\mathbf{M}}_k(j) = (N_k)^{-1}\sum_{i=1}^{N_k}\mathbf{X}_i^k(j) = \mathbf{U}_k(j), \tag{22}$$

$$\hat{\mathbf{\Sigma}}_k^j = (N_k)^{-1}\sum_{i=1}^{N_k}(\mathbf{X}_i^k(j) - \mathbf{U}_k(j))(\mathbf{X}_i^k(j) - \mathbf{U}_k(j))^{\mathrm{T}}, \tag{23}$$

where $\mathbf{X}_i^k(j)$ is the jth column of the ith sample matrix of class C_k as defined previously.

Then, based on equalities (17)–(23), the following theorem first gives the sufficient conditions when 2D-LDA would be Bayes optimal for two-class classification problem. Its proof can be found in Appendix B.

Theorem 2. *For two-class classification problem*, 2*D-LDA in terms of equality* (3) *is Bayes optimal if the following conditions hold*:

(1) *Columns with different indexes of image matrices are independent*, *i.e.*, *equality* (17).

(2) *Columns with the same index of image matrices within each class are normally distributed*, *i.e.*, *equality* (20), *and the covariance matrices are equal as follows*:

$$\hat{\mathbf{\Sigma}}_{k_1}^{j_1} = \hat{\mathbf{\Sigma}}_{k_2}^{j_2} = \tilde{\mathbf{S}}_w, \quad \forall j_1\neq j_2, k_1\neq k_2,$$
$$\tilde{\mathbf{S}}_w = \sum_{k=1}^{2}\sum_{j=1}^{col} P(C_k, j)\left\{(N_k)^{-1}\sum_{i=1}^{N_k}(\mathbf{X}_i^k(j) - \mathbf{U}_k(j))(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))^{\mathrm{T}}\right\},$$
$$P(C_k, j) = N_k\cdot(N\cdot col)^{-1}. \tag{24}$$

(3) *Differences between any two columns with the same index of two class mean matrices are equal except some scalar*

scaling, i.e., there exist $s_i \neq 0$, $i = 1, \ldots, col$, *such that*

$$\Delta\mathbf{U} = s_i(\mathbf{U}_1(i) - \mathbf{U}_2(i)) = s_j(\mathbf{U}_1(j) - \mathbf{U}_2(j)), \quad \forall i \neq j, i, j = 1, \ldots, col. \tag{25}$$

Those sufficient conditions could help understand some findings presented. It is because if condition (1) is satisfied then it is true why 2D-LDA in terms of equality (3) eliminates the relations between different columns, and if conditions (2)–(3) are valid it would be interpretable that why 2D-LDA estimates its between-class scatter matrix by averaging the between-class scatter matrices over all column indexes and also model the within-class scatter matrix by averaging the within-class scatter matrices over all column indexes.

Being Bayes optimal, 2D-LDA presented above, however, requires more conditions than 1D-LDA. Then, what are the differences between 1D-LDA and 2D-LDA when those conditions in Theorem 2 are satisfied or not satisfied? We finally give a discussion below. First, we note that for any given $\mathbf{X} = [\mathbf{X}(1), \ldots, \mathbf{X}(col)]$, its vector form is $\mathbf{x} = [\mathbf{X}(1)^{\mathrm{T}}, \ldots, \mathbf{X}(col)^{\mathrm{T}}]^{\mathrm{T}}$. Then it is true that

$$p(\mathbf{X}) = p([\mathbf{X}(1), \ldots, \mathbf{X}(col)]) = p([\mathbf{X}(1)^{\mathrm{T}}, \ldots, \mathbf{X}(col)^{\mathrm{T}}]) = p([\mathbf{X}(1)^{\mathrm{T}}, \ldots, \mathbf{X}(col)^{\mathrm{T}}]^{\mathrm{T}}) = p(\mathbf{x}), \tag{26}$$

$$p(\mathbf{X}|C_k) = p(\mathbf{x}|C_k), \tag{27}$$

$$p(C_k|\mathbf{X}) = p(C_k|\mathbf{x}). \tag{28}$$

Hence the declaration "$\mathbf{X} \in C_1$ if and only if $p(C_1|\mathbf{X}) > p(C_2|\mathbf{X})$, else $\mathbf{X} \in C_2$" is equivalent to the one "$\mathbf{X} \in C_1$ if and only if $p(C_1|\mathbf{x}) > p(C_2|\mathbf{x})$, else $\mathbf{X} \in C_2$." Therefore for two-class classification problem, we could have the following:

(1) If those sufficient conditions (1)–(3) in Theorem 2 are satisfied, both 1D-LDA and 2D-LDA are Bayes optimal. The vector-form sample $\mathbf{x}=[\mathbf{X}(1)^{\mathrm{T}}, \ldots, \mathbf{X}(col)^{\mathrm{T}}]^{\mathrm{T}}$ is then normally distributed with equal covariance matrix within each class under conditions (1)–(2), and the covariance matrix of $\mathbf{x}$ within class C_k is indicated by equality (29) below under condition (1):

$$\begin{aligned}
&\mathbf{E}[(\mathbf{x} - \mathbf{E}[\mathbf{x}|C_k])(\mathbf{x} - \mathbf{E}[\mathbf{x}|C_k])^{\mathrm{T}}|C_k] \\
&= \mathbf{E}\left[\left.\begin{pmatrix} \mathbf{X}(1) - \mathbf{E}[\mathbf{X}(1)|C_k] \\ \vdots \\ \mathbf{X}(col) - \mathbf{E}[\mathbf{X}(col)|C_k] \end{pmatrix}\begin{pmatrix} \mathbf{X}(1) - \mathbf{E}[\mathbf{X}(1)|C_k] \\ \vdots \\ \mathbf{X}(col) - \mathbf{E}[\mathbf{X}(col)|C_k] \end{pmatrix}^{\mathrm{T}}\right| C_k\right] \\
&= \begin{bmatrix} \mathbf{E}[(\mathbf{X}(1) - \mathbf{E}[\mathbf{X}(1)|C_k])(\mathbf{X}(1) - \mathbf{E}[\mathbf{X}(1)|C_k])^{\mathrm{T}}|C_k] & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{E}[(\mathbf{X}(col) - \mathbf{E}[\mathbf{X}(col)|C_k])(\mathbf{X}(col) - \mathbf{E}[\mathbf{X}(col)|C_k])^{\mathrm{T}}|C_k] \end{bmatrix},
\end{aligned} \tag{29}$$

where the estimations of $\mathbf{E}[(\mathbf{X}(j) - \mathbf{E}[\mathbf{X}(j)|C_k])(\mathbf{X}(j) - \mathbf{E}[\mathbf{X}(j)|C_k])^{\mathrm{T}}|C_k]$, $j=1, \ldots, col$, $k=1, 2$ are equal under condition (2).

(2) If only conditions (1)–(2) are satisfied, 1D-LDA could be Bayes optimal, while there is no guarantee for 2D-LDA being Bayes optimal. Hence one could recall equality (16) which indicates that why 1D-LDA is better than 2D-LDA in such case, i.e., condition (1).

(3) If $\mathbf{X}(1), \ldots, \mathbf{X}(col)$ are not independent, then 2D-LDA in terms of equality (3) loses discriminative information in the covariance information between different columns of an image. Generally speaking, condition (1) is not required for 1D-LDA to be Bayes optimal.

(4) If conditions (2)–(3) are not satisfied, then the heteroscedastic problem in 2D-LDA discussed cannot be avoided.

(5) Finally, we see that if vector sample $\mathbf{x} = [\mathbf{X}(1)^{\mathrm{T}}, \ldots, \mathbf{X}(col)^{\mathrm{T}}]^{\mathrm{T}}$ is normally distributed with equal class covariance matrices, then 1D-LDA is Bayes optimal, but those conditions (1)–(3)for 2D-LDA cannot be implied in such case.

3.4. Why is 2D-LDA sometimes superior?

The above analysis on 2D-LDA is based on the equality (3). Actually some similar conclusions could also be obtained for its variations. First, we see that if the image matrices are first transposed, equality (4) would become equality (3). Even though B-2D-LDA has combined both approaches, however, it is hard to obtain a closed form solution. So far there are at least two ways to find a practical solution of B-2D-LDA. One way is to drive an iterative algorithm that finds the optimal value for $\mathbf{w}_{l\text{-}opt}^{2d}$ while fixing $\mathbf{w}_{r\text{-}opt}^{2d}$ and finds the optimal value for $\mathbf{w}_{r\text{-}opt}^{2d}$ while fixing $\mathbf{w}_{l\text{-}opt}^{2d}$ [23,54]. Another way is to calculate them independently and then combine them [28]. Hence the potential drawbacks of 2D-LDA discussed above are embedded in each process of computation of B-2D-LDA.

However, why has 2D-LDA been recently reported superior to some 1D-LDA based algorithms experimentally? The reasons may be the following:

(1) The dimensionality of the optimal feature $\mathbf{w}_{opt}^{2d}$ extracted by 2D-LDA is much smaller than the one $\mathbf{w}_{opt}$ extracted by 1D-LDA, while the number of samples learned for $\mathbf{w}_{opt}^{2d}$ is actually much larger than the one for $\mathbf{w}_{opt}$, because for 2D-LDA each column or each row of an image is a training sample, while for 1D-LDA only the whole image is a training sample. Therefore, the number of parameters estimated for $\mathbf{w}_{opt}^{2d}$ is much less than the one for $\mathbf{w}_{opt}$ and the bias of the estimation of $\mathbf{w}_{opt}^{2d}$ could be smaller than the estimation of $\mathbf{w}_{opt}$.

(2) 1D-LDA is always confronted with the singularity problem. For 1D-LDA, the strategy to overcome such problem

is crucially important. So far some standard approaches are proposed [8,35–38,42–46]. It is known that most of the dimension reduction techniques for 1D-LDA would lose discriminant information, such as Fisherface and nullspace LDA. In contrast, 2D-LDA would always avoid the singularity problem. However, some well-known standard approaches of 1D-LDA, such as nullspace LDA and regularized LDA, have been presented to be effective and powerful in practice, but previous experimental results have rarely reported the comparison of 2D-LDA with them, especially regularized LDA which is almost a pure LDA except the additional regularization term. Thus this paper would like to include them for comparison.

(3) The data set selected for comparison is important. Moreover in the experiment, we will find that the final classifier is indeed an impact in evaluating the performances of 1D-LDA and 2D-LDA. However, it is also not suggested before.

4. 1 D-LDA vs. 2D-LDA: experimental comparison

Besides theoretical comparison, a comprehensive experimental comparison between 1D-LDA and 2D-LDA is also performed here. The main goal is to compare them under the case when the number of training samples for each class is limited or when the number of discriminant features used is small. Some existing views will be broken. Experimental results are reported on FERET [55] and CMU [56] databases. As either 2D-LDA or 1D-LDA is actually used for discriminant feature extraction, a final classifier is employed for classification in the feature space. Two such classifiers, namely nearest neighbor classifier (NNC) and nearest class mean classifier (NCMC) are employed to evaluate the performances. They are always popularly used for evaluation of the LDA-based algorithms and it will be shown that the final classifier would have an impact on the performances of some algorithms. Note that in almost all published papers regarding 2D-LDA only NNC is selected as the final classifier [21–23,27,29,30].

We compare some standard 1D-LDA based algorithms with some standard 2D-LDA based algorithms. The compared 1D-LDA based algorithms involve Fisherface, nullspace LDA and regularized LDA. For comparison, they are renamed as "1D-LDA, Fisherface", "1D-LDA, nullspace LDA" and "1D-LDA, regularized LDA". Regularized LDA is implemented by equality (2) with $\lambda = 0.005$. For 2D-LDA, we have implemented its three standard algorithms, i.e., equalities (3)–(5). For comparison, they are also renamed as "unilateral 2D-LDA, left" (equality (3)), "unilateral 2D-LDA, right (equality (4)), and "bilateral 2D-LDA" (equality (5)), where the number of iteration in "bilateral 2D-LDA" is set to be 10. Noting that regularized LDA is almost a pure 1D-LDA except the regularization term added to the within-class scatter matrix, hence it is valuable to take it into comparison.

Table 1
Brief descriptions of databases and subsets used

Database/subset	Number of persons	Number of faces (per person)	Database/subset size	Image size
FERET	255	4	1020	92×112
CMU-NearFrontalPose-Expression	68	15	1020	60×80
CMU-Illumination-Frontal	68	43	2924	60×80
CMU-11-Poses	68	11	748	60×80

4.1. Introduction to databases and subsets

A large subset of FERET [55] is established by extracting images from four different sets, namely Fa, Fb, Fc and

Fig. 1. Illustrations of some face images (images are resized to show): (a) FERET; (b) CMU-Illumination-Frontal; (c) CMU-NearFrontalPose-Expression; (d) CMU-11-Poses.

duplicate. It consists of 255 persons, and for each individual there are four face images undergoing expression variation, illumination variation, age variation, etc.

Three subsets of CMU PIE [56] are also established, called "CMU-NearFrontalPose-Expression", "CMU-Illumination-Frontal" and "CMU-11-Poses". The subset "CMU-Near-FrontalPose-Expression" is established by selecting images under natural illumination for all persons from the frontal view, $\frac{1}{4}$ left\right profile and below\above in frontal view. For each view, there are three different expressions, namely natural expression, smiling and blinking [56]. Hence there are 15 face images for each object. The subset "CMU-Illumination-frontal" consists of images with all illumination variations in Frontal view under the background light off and on. The subset "CMU-11-Poses" consists of images across 11 different poses of each person, including $\frac{3}{4}$ right profile, half right profile, $\frac{1}{4}$ right profile, frontal view, $\frac{1}{4}$ left profile, half left profile, $\frac{3}{4}$ left profile, below in frontal view, above in frontal view and two surveillance views, and all images are under natural illumination and natural expression.

The data sets used are briefly summarized in Table 1 and some face images are illustrated in Fig. 1. Note that all images are linearly stretched to full range of pixel values of [0, 1].

Table 2
Range of the number of training samples for each class

Database	Range
FERET	[2 : 1 : 3]
CMU-NearFrontalPose-Expression	[2 : 1 : 8][a]
CMU-Illumination-Frontal	[2 : 1 : 8]
CMU-11-Poses	[2 : 1 : 8]

[a][2 : 1 : 8] means the number of training samples for each class ranges from 2 to 8 with step 1.

4.2. Comparison

For each data set, the comparisons involve two parts. In the first part, the number of training samples for each class is fixed, and the average recognition rates of an algorithm with respect to different numbers of discriminant features are presented. Based on these results, we then illustrate the best average recognition rates of an algorithm with respect to different numbers of training samples for each class in the second part. Results are reported based on NCMC and NNC, respectively. Additionally, for an algorithm tested on a data set, if the number of discriminant features used is fixed and there are Num_T training samples for each class, then the test procedure will be repeated 10 times. For each time, Num_T samples are randomly selected from each class to establish the training set and the rest are for testing. The average recognition rate is got finally.

4.2.1. Recognition rate vs. number of discriminant features

This section first presents the experimental results to show how the average recognition rates of the LDA-based algorithms change depending on the number of extracted discriminant features used when the number of training samples for each class is fixed. In Table 2, the range of the variation of the number of training samples for each class is indicated. Since the experimental analysis would like to focus on comparing different LDA algorithms in the small sample size case that is when the training sample size for each class is limited, so the average recognition rates are not reported when the number of training samples for each class is more than 8 over three CMU subsets. Solving the small sample size problem is a strong motivation for many proposed LDA algorithms in the past several years, including the compared ones in this paper.

For an algorithm, suppose its maximum number of discriminant features is Num_AF. Then its all features are ordered

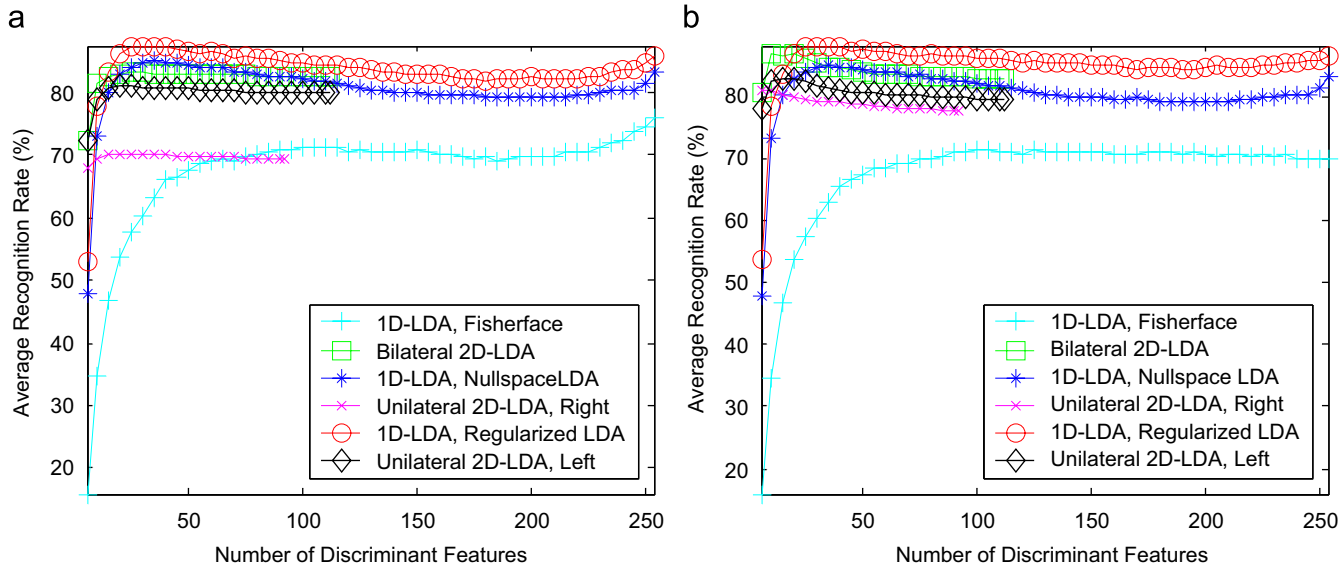

Fig. 2. Recognition rate vs. number of discriminant features on FERET; training number is three for each class. (a) Final classifier: NCMC. (b) Final classifier: NNC.

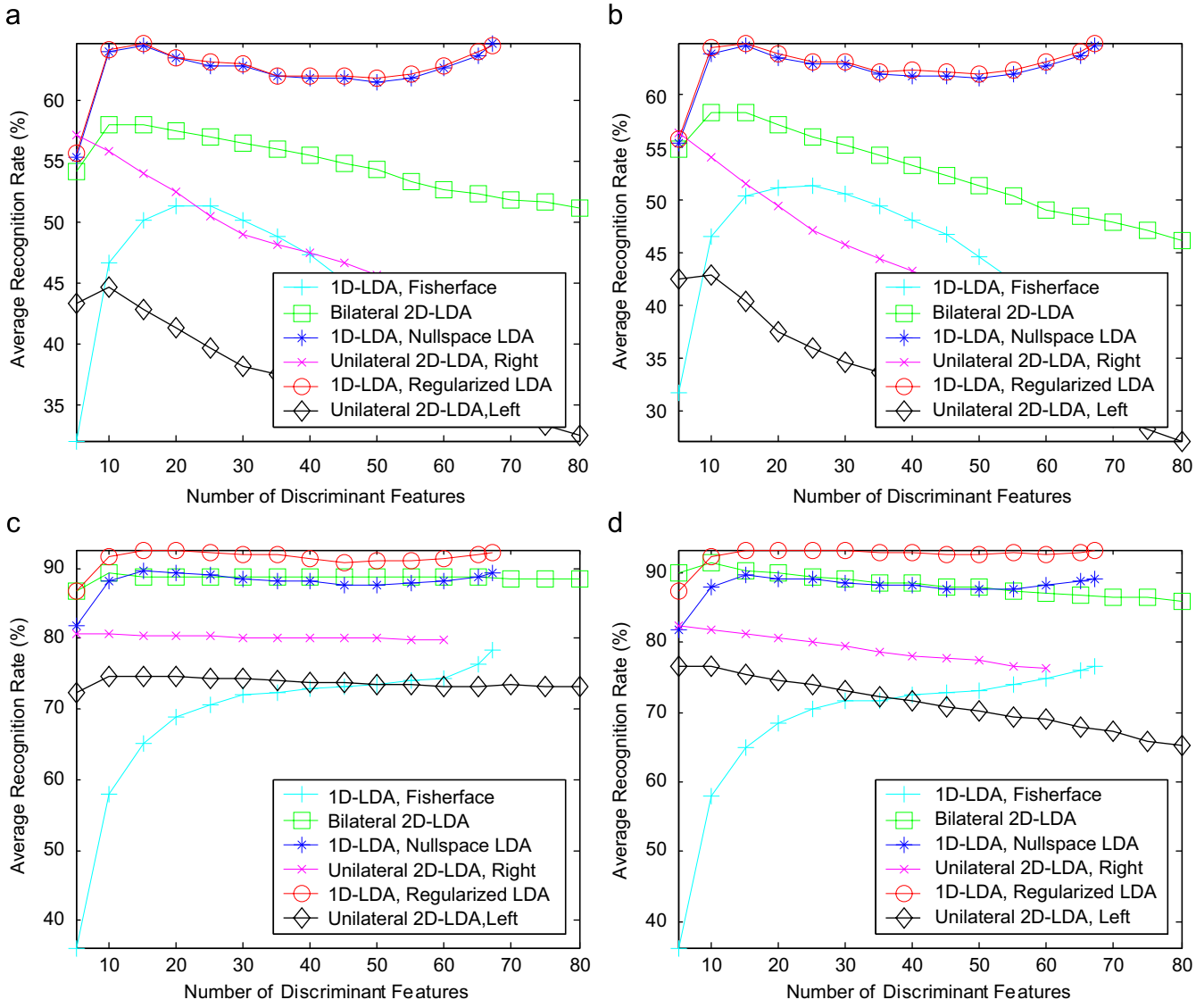

Fig. 3. Recognition rate vs. number of discriminant features on "CMU-NearFrontalPose-Expression"; training number is two for each class in (a)–(b) and training number is seven for each class in (c)–(d). (a) Final classifier: NCMC. (b) Final classifier: NNC. (c) Final classifier: NCMC. (d) Final classifier: NNC.

according to their corresponding eigenvalues in a descendant order, since the eigenvalue of each feature could be treated as a measurement of the discriminative ability. Finally, the top Num_F features are selected to evaluate the recognition performance, where we would let $Num\_F = 5, 10, 15, 20, \ldots, Num\_AF$. Additionally, the scheme for "bilateral 2D-LDA" is explained as follows. "bilateral 2D-LDA" has bilateral projections, while the maximum numbers of features with respect to two different side projections are always different. Hence, if there are Num_F features selected for "bilateral 2D-LDA", it means the top Num_F features are selected for both projections, respectively. If the value Num_F has exceeded the maximum number of features of one of the projections, then all features of that projection would be used.

Due to the limited length of the paper, only some figures describe the experiment results could be illustrated. For FERET database, we present the results when the number of training samples for each class is three (Fig. 2); for "CMU-NearFrontalPose-Expression" and "CMU-11-Poses", it is 2 and 7 in Figs. 3 and 5; for "CMU-Illumination-Frontal" it is 3 and 7 in Fig. 4. The sample size of FERET is limited so we only present the case when the number of training samples for each class is three; for "CMU-Illumination-Frontal" the result when the number of training samples for each class is 3 rather than 2 is presented, because the performance of Fisherface increases notably as observed later in Fig. 7 when NCMC is used. The best average recognition rates with respect to different numbers of training samples for each class will be totally reported in the next section.

From the experimental results above, it could observed that the 2D-LDA based algorithms always achieve their best performances when the number of discriminant features is retained appropriately small while the performances of them would sometimes degrade if more features are used. Interestingly, the 1D-LDA based algorithms may also achieve their best performances sometimes when an appropriately small set of features

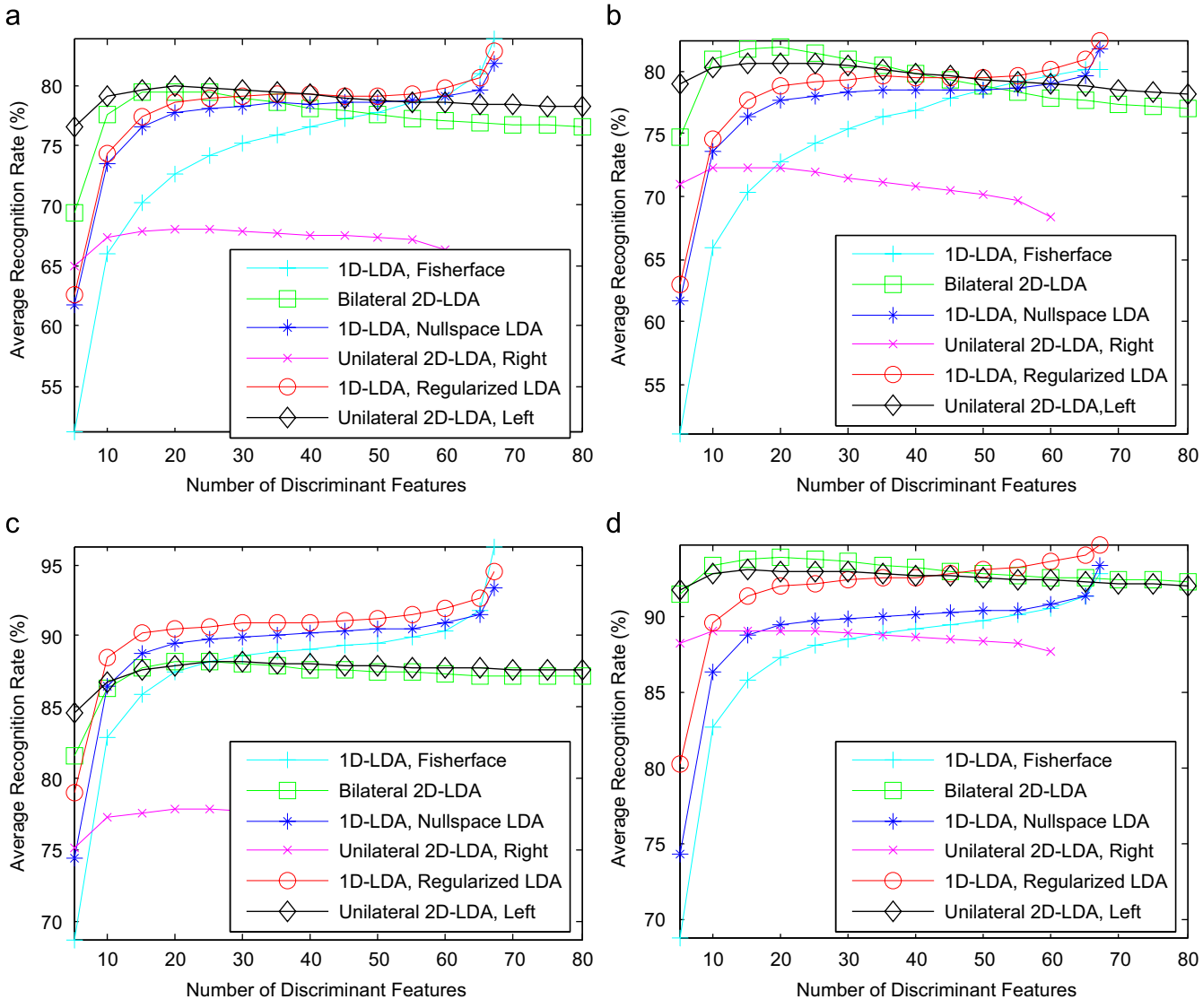

Fig. 4. Recognition rate vs. number of discriminant features on "CMU-Illumination-Frontal"; training number is three for each class in (a)–(b) and training number is seven for each class in (c)–(d). (a) Final classifier: NCMC. (b) Final classifier: NNC. (c) Final classifier: NCMC. (d) Final classifier: NNC.

is retained. However, sometimes their performances would first descend and then ascend as more features are used. Such scenario could be obviously observed in Fig. 3(a)–(b) and Fig. 5. A recent developed theory on LDA by Martínez and Zhu has told the fact that not all discriminant features are good for classification [57]. Hence a small set of features would sometimes get its best accuracy. Of course it is not always the case, since the 2D-LDA based algorithms do not degrade too much in Fig. 3(c)–(d) when more features are used and the 1D-LDA based algorithms perform better and better in Fig. 4 when more features are used. However, it could be found that if all features of the 1D-LDA based algorithms are used, the performances are always almost the same as their best ones acquired, but it is not always the case for the 2D-LDA based algorithms. Therefore, the experimental results indicate how to select the proper number of features would potentially be a more serious problem for the 2D-LDA based algorithms than that for the 1D-LDA based algorithms.

The experiments have also broken the existing viewpoint that 2D-LDA could always achieve better performance than 1D-LDA when only fewer discriminant features are used [21,22], since it is also found that regularized LDA and nullspace LDA could achieve their best performances and perform better than the 2D-LDA based algorithms on data sets FERET (Fig. 2), "CMU-NearFrontalPose-Expression" (Fig. 3) and "CMU-11-Poses" (Fig. 5(c)–(d)) when fewer features are used. Note that even Fisherface could perform better than some 2D-LDA based algorithms if a little more discriminant features are employed sometimes.

4.2.2. *Recognition rate vs. number of training samples*

This section shows how the best average recognition rate of an algorithm changes depending on the number of training samples for each class. Except FERET database, all experimental results are presented in figures. In all tables and figures, the best average recognition rates for fixed number of training samples

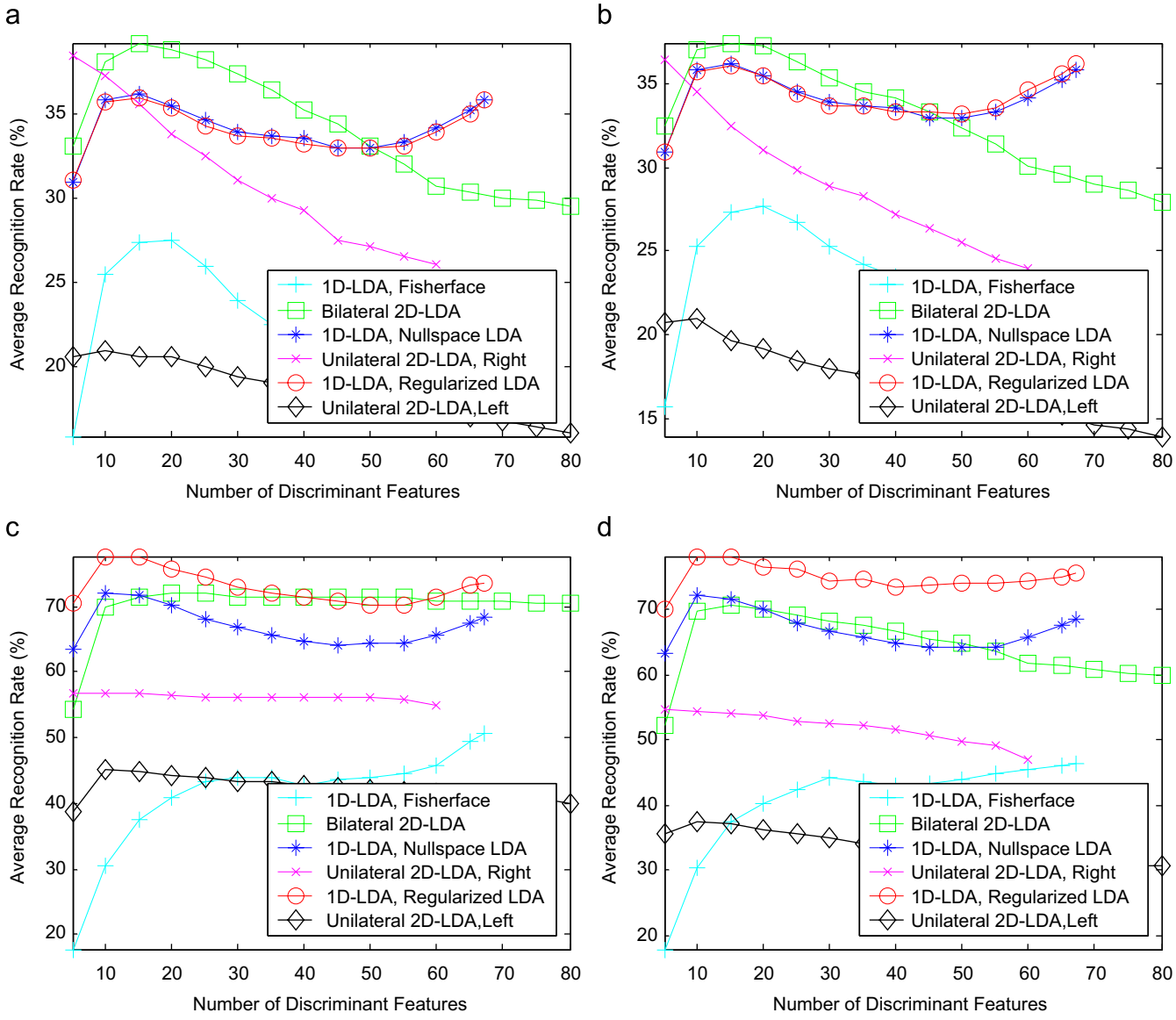

Fig. 5. Recognition rate vs. number of discriminant features on "CMU-11-Poses"; training number is two for each class in (a)–(b) and training number is seven for each class in (c)–(d). (a) Final classifier: NCMC. (b) Final classifier: NNC. (c) Final classifier: NCMC. (d) Final classifier: NNC.

for each class are reported. For each algorithm, the best average recognition rate is the highest one among the corresponding average recognition rates with respect to different numbers of discriminant features, which are reported in the last section. It would be a fair comparison, as the number of discriminant features used has an obvious impact on the performance of an algorithm as observed in the first part.

From the experiments, it could be observed that the 2D-LDA based algorithms almost always perform better than Fisherface except the experiment on "CMU-Illumination-Frontal" (Fig. 7) where Fisherface performs the best by using NCMC there when the number of training samples for each class is larger than three. Though it is known that Fisherface loses discriminant information [35,36,42,58], however it has also been known that Fisherface was first proposed to handle various illuminations [8] for face recognition, while images in "CMU-Illumination-Frontal" are just corrupted by illuminations and no other variations exist there. The performance of Fisherface would dramatically reduce if other variations, such as pose or expression, are involved. However, we observe that regularized LDA and nullspace LDA always obtain superior performances than the 2D-LDA based algorithms on some data sets. This could be obviously found from the experiments on the data sets FERET (Table 3),"CMU-NearFrontalPose-Expression" (Fig. 6) and "CMU-11-Poses" (Fig. 8). Note that Nullspace LDA would perform the same no matter NCMC or NNC is used. It is because the projection on the nullspace of the within-class scatter matrix has already transformed each training sample to its class center [37]. Other than nullspace LDA, the superiority of regularized LDA is more notable no matter which final classifier is used. It may be because regularized LDA only adds a small regularization to the within-class scatter matrix and it is almost a purely naive Fisher's LDA algorithm while nullspace LDA still discards some discriminant information [58].

Table 3
Best average recognition rate on FERET

Final classifier	NCMC (%)		NNC (%)	
Number of training samples for each class	2	3	2	3
1D-LDA, Fisherface	63.51	76.20	63.59	71.61
1D-LDA, nullspace LDA	76.10	85.10	76.10	85.10
1D-LDA, regularized LDA	77.35	**87.53**	77.37	**88.27**
Bilateral 2D-LDA	75.84	83.33	76.29	87.14
Unilateral 2D-LDA, right	65.63	70.12	68.78	81.18
Unilateral 2D-LDA, left	73.51	81.29	72.51	83.10

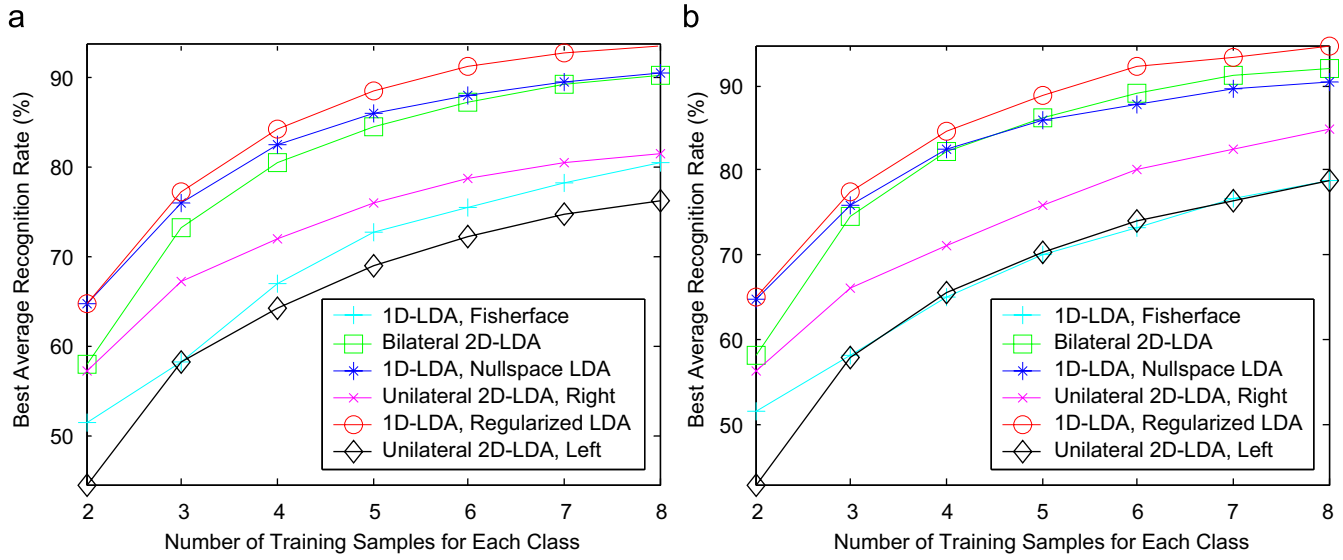

Fig. 6. Recognition rate vs. number of training samples on "CMU-NearFrontalPose-Expression". (a) Final classifier: NCMC. (b) Final Classifier: NNC.

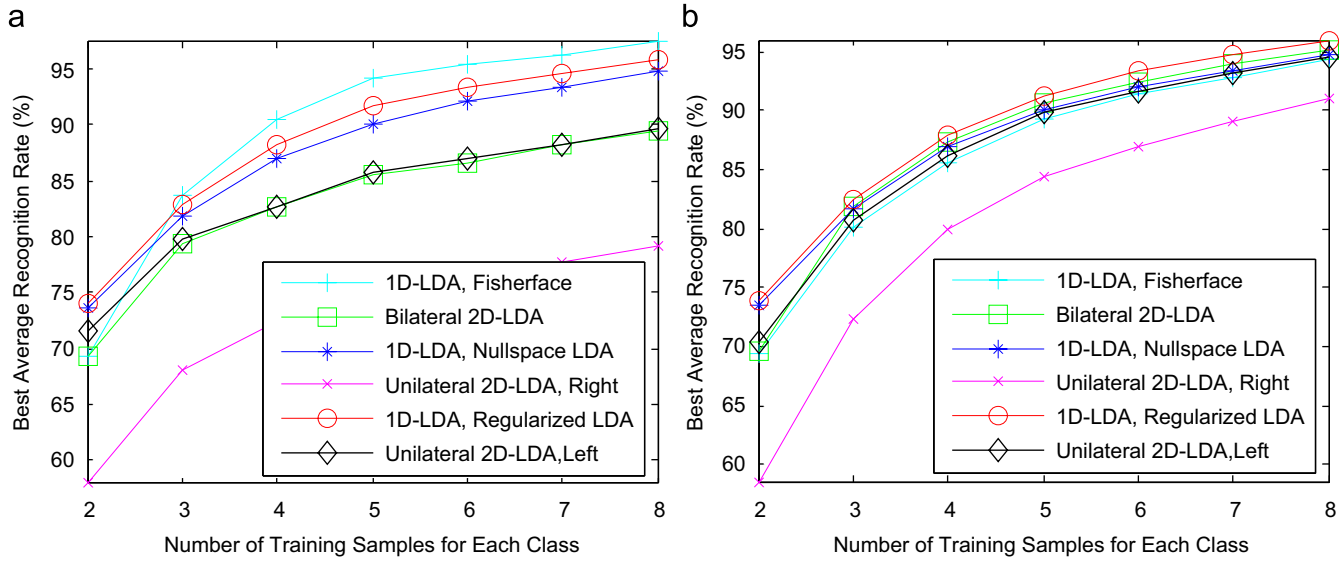

Fig. 7. Recognition rate vs. number of training samples on "CMU-Illumination-Frontal". (a) Final classifier: NCMC; (b) Final classifier: NNC.

Actually, some 2D-LDA based algorithms do not perform well over some challenging data sets. For instance, both "unilateral 2D-LDA, left" and "unilateral 2D-LDA, right" do not have satisfied performances on "CMU-NearFrontalPose-Expression" and "CMU-11-Poses" no matter if NCMC or NNC is used, and "unilateral 2D-LDA, right" does not perform well over "CMU-Illumination-Frontal" using NCMC. However, "bilateral 2D-LDA" would perform more stable. It outperforms some 1D-LDA based algorithms on "CMU-Illumination-Frontal" data set, and it performs the best especially when only two

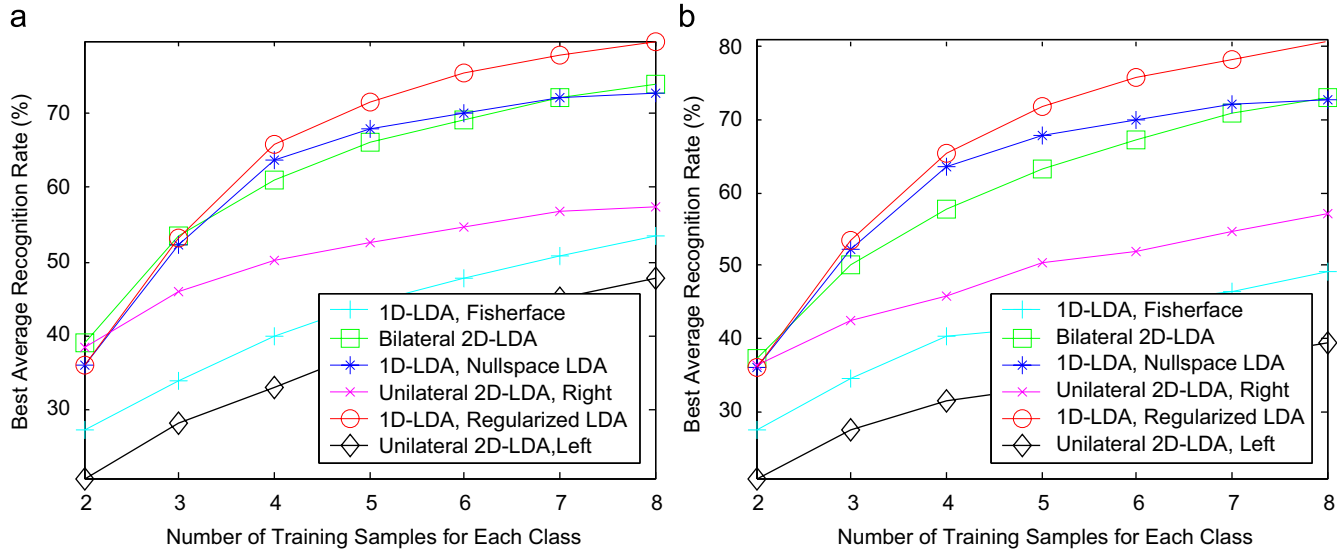

Fig. 8. Recognition rate vs. number of training samples on "CMU-11-Poses". (a) Final classifier: NCMC. (b) Final classifier: NNC.

samples for each class are used for training as shown in Figs. 5(a)–(b) and 8.

From the experimental results, it is found that the performance of 2D-LDA is sometimes sensitive to the final classifier. As indicated by Table 3 and Figs. 6–8, most 2D-LDA based algorithms could improve their recognition performances obviously if NNC rather than NCMC is used. In contrast, the 1D-LDA based algorithms are less sensitive. For Fisherface, NCMC may be more preferred, but for regularized LDA, using NNC would be a little better. However, it does not mean the 2D-LDA based algorithms would outperform the 1D-LDA based algorithms if NNC is employed.

Hence there is no convinced evidence that the 2D-LDA based algorithms could always outperform the 1D-LDA based algorithms if the number of training samples for each class is small, and it also breaks the existing view on this issue [27,28].

In fact, some experimental results above also agree with some published results [21–23,28] that some 2D-LDA based algorithms like "bilateral 2D-LDA" and "unilateral 2D-LDA, right" are reported to always get superior performance to Fisherface. However, it is not always true due to the experimental results reported on "CMU-Illumination-Frontal", in which images are only purely undergoing illumination. Compared with the published papers, more extensive comparisons have been provided between 1D-LDA and 2D-LDA, by comparing the performances of them depending on the number of discriminant features used and the number of training samples for each class. Moreover, some existing views are broken. In addition, we find that just a small regularization term could thoroughly enhance the performance of 1D-LDA like regularized LDA. The comparison between regularized LDA and the 2D-LDA based algorithms has not been reported before.

5. Summarization

In order to investigate when vector-based LDA would be better, we present theoretical and experimental analyses between 1D-LDA and 2D-LDA. The findings are briefly listed below:

(1) 2D-LDA would also be confronted with the heteroscedastic problem, and it would be more serious for 2D-LDA than 1D-LDA.
(2) Relationship between 1D-LDA and 2D-LDA are explored and modeled in equalities. It gives a new way to find 2D-LDA actually loses the covariance information between different local structures, while 1D-LDA could preserve such information. It is further found that the Fisher score of 1D-LDA is higher than the one gained by 2D-LDA in the extreme case.
(3) For two-class classification problem, the sufficient conditions for 2D-LDA being Bayes optimal are given. Discussions between 1D-LDA and 2D-LDA are also presented when those sufficient conditions are satisfied or not, supporting the other findings in this paper.
(4) Existing views are broken in the experiment and it is found there is no convinced evidence that 2D-LDA would always outperform 1D-LDA when the number of training samples for each class is small or when the number of discriminant features used is small. Sometimes 1D-LDA, especially regularized LDA, performs better. Besides the choice of final classifier, it is also found that selecting the appropriate number of features would be a more serious problem in 2D-LDA than that in 1D-LDA.

However, it is known that 2D-LDA could always avoid the singularity problem of within-class scatter matrix while 1D-LDA would be always confronted with it in practice. Moreover, for 2D-LDA each column or each row of an image could be treated as a training sample while only the whole image could be a sample for 1D-LDA. Hence, from the bias estimation point of view, 2D-LDA might be more stable since more samples are actually used for learning.

Finally, it is stressed that this paper does not aim to declare which algorithm is the best. We investigate into the question by presenting a fair comparison between 1D-LDA and 2D-LDA in both theoretical and experimental sense. The goal of the extensive comparisons is to explore the properties of 2D-LDA, present its disadvantages and some inherent problems, and find when 1D-LDA would be better. Even though some 2D-LDA based algorithms do not perform as well as some standard 1D-LDA based algorithms in the experiments, it still does not mean 2D-LDA is not effective sometimes.

In conclusion, our findings indicate that using the matrix-based feature extraction technique would not always result in a better performance than using the traditional vector-form representation. The traditional vector-form representation is still useful.

Acknowledgements

This project was supported by the National Natural Science Foundation of China (60373082), 973 Program (2006CB303104), the Key (Key grant) Project of Chinese Ministry of Education (105134) and NSF of Guangdong (06023194). The authors would also like to thank the reviewers for their constructive advice.

Appendix A. Proof of Lemma 1

As indicated at the beginning of Section 3, we note that $\mathbf{x}_i^k = [\mathbf{X}_i^k(1)^{\mathrm{T}}, \ldots, \mathbf{X}_i^k(col)^{\mathrm{T}}]^{\mathrm{T}}$. Then we have

$$
\begin{aligned}
&\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w} \\
&= \sum_{k=1}^{L} \frac{N_k}{N} \mathbf{w}^{\mathrm{T}}(\mathbf{u}_k - \mathbf{u})(\mathbf{u}_k - \mathbf{u})^{\mathrm{T}}\mathbf{w} \\
&= \sum_{k=1}^{L} \frac{N_k}{N} [\widehat{\mathbf{w}}_1^{\mathrm{T}}, \ldots, \widehat{\mathbf{w}}_{col}^{\mathrm{T}}] \begin{bmatrix} \mathbf{U}_k(1) - \mathbf{U}(1) \\ \vdots \\ \mathbf{U}_k(col) - \mathbf{U}(col) \end{bmatrix} \\
&\quad \times [(\mathbf{U}_k(1) - \mathbf{U}(1))^{\mathrm{T}}, \ldots, (\mathbf{U}_k(col) - \mathbf{U}(col))^{\mathrm{T}}] \begin{bmatrix} \widehat{\mathbf{w}}_1 \\ \vdots \\ \widehat{\mathbf{w}}_{col} \end{bmatrix} \\
&= \sum_{k=1}^{L} \frac{N_k}{N} \left[\sum_{j=1}^{col} \widehat{\mathbf{w}}_j^{\mathrm{T}}(\mathbf{U}_k(j) - \mathbf{U}(j)) \right] \\
&\quad \times \left[\sum_{j=1}^{col} (\mathbf{U}_k(j) - \mathbf{U}(j))^{\mathrm{T}} \widehat{\mathbf{w}}_j \right] \\
&= \sum_{k=1}^{L} \frac{N_k}{N} \sum_{j=1}^{col} \widehat{\mathbf{w}}_j^{\mathrm{T}}(\mathbf{U}_k(j) - \mathbf{U}(j))(\mathbf{U}_k(j) - \mathbf{U}(j))^{\mathrm{T}} \widehat{\mathbf{w}}_j \\
&\quad + \sum_{k=1}^{L} \frac{N_k}{N} \sum_{j=1, h=1, j\neq h}^{col} \widehat{\mathbf{w}}_j^{\mathrm{T}}(\mathbf{U}_k(j) - \mathbf{U}(j))(\mathbf{U}_k(h) \\
&\quad - \mathbf{U}(h))^{\mathrm{T}} \widehat{\mathbf{w}}_h, \qquad \text{(A.1)}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} \\
&= \frac{1}{N} \sum_{k=1}^{L} \sum_{i=1}^{N_k} \mathbf{w}^{\mathrm{T}}(\mathbf{x}_i^k - \mathbf{u}_k)(\mathbf{x}_i^k - \mathbf{u}_k)^{\mathrm{T}}\mathbf{w} \\
&= \frac{1}{N} \sum_{k=1}^{L} \sum_{i=1}^{N_k} [\widehat{\mathbf{w}}_1^{\mathrm{T}}, \ldots, \widehat{\mathbf{w}}_{col}^{\mathrm{T}}] \begin{bmatrix} \mathbf{X}_i^k(1) - \mathbf{U}_k(1) \\ \vdots \\ \mathbf{X}_i^k(col) - \mathbf{U}_k(col) \end{bmatrix} \\
&\quad \times [(\mathbf{X}_i^k(1) - \mathbf{U}_k(1))^{\mathrm{T}}, \ldots, (\mathbf{X}_i^k(col) - \mathbf{U}_k(col))^{\mathrm{T}}] \\
&\quad \times \begin{bmatrix} \widehat{\mathbf{w}}_1 \\ \vdots \\ \widehat{\mathbf{w}}_{col} \end{bmatrix} \\
&= \frac{1}{N} \sum_{k=1}^{L} \sum_{i=1}^{N_k} \left[\sum_{j=1}^{col} \widehat{\mathbf{w}}_j^{\mathrm{T}}(\mathbf{X}_i^k(j) - \mathbf{U}_k(j)) \right] \\
&\quad \times \left[\sum_{j=1}^{col} (\mathbf{X}_i^k(j) - \mathbf{U}_k(j))^{\mathrm{T}} \widehat{\mathbf{w}}_j \right] \\
&= \frac{1}{N} \sum_{k=1}^{L} \sum_{i=1}^{N_k} \sum_{j=1}^{col} \widehat{\mathbf{w}}_j^{\mathrm{T}}(\mathbf{X}_i^k(j) - \mathbf{U}_k(j))(\mathbf{X}_i^k(j) - \mathbf{U}_k(j))^{\mathrm{T}} \widehat{\mathbf{w}}_j \\
&\quad + \frac{1}{N} \sum_{k=1}^{L} \sum_{i=1}^{N_k} \sum_{j=1, h=1, j\neq h}^{col} \widehat{\mathbf{w}}_j^{\mathrm{T}}(\mathbf{X}_i^k(j) - \mathbf{U}_k(j))(\mathbf{X}_i^k(h) \\
&\quad - \mathbf{U}_k(h))^{\mathrm{T}} \widehat{\mathbf{w}}_h. \qquad \text{(A.2)}
\end{aligned}
$$

Using equalities (6)–(7) and equalities (8), (11), the lemma is then proved.

Appendix B. Proof of Theorem 2

Based on equalities (21)–(23), substituting the estimates of the means and the covariance matrices and eliminating the ineffective ingredients that do not affect the classification result in formula (21) would yield the following Bayes classifier:

$$
\begin{aligned}
g_k(\mathbf{X}) = \sum_{j=1}^{col} \Bigg\{ &-\frac{1}{2} \log |\hat{\boldsymbol{\Sigma}}_k^j| - \frac{1}{2}(\mathbf{X}(j) - \mathbf{U}_k(j))^{\mathrm{T}} (\hat{\boldsymbol{\Sigma}}_k^j)^{-1} (\mathbf{X}(j) \\
&- \mathbf{U}_k(j)) \Bigg\} + \log(P(C_k)). \qquad \text{(B.1)}
\end{aligned}
$$

Under the condition (2) in the theorem, $\hat{\boldsymbol{\Sigma}}_k^j$ are equal. We hence further have

$$
\begin{aligned}
g_k(\mathbf{X}) &= \sum_{j=1}^{col} \Bigg\{ -\frac{1}{2} \log |\tilde{\mathbf{S}}_w| - \frac{1}{2}(\mathbf{X}(j) - \mathbf{U}_k(j))^{\mathrm{T}} (\tilde{\mathbf{S}}_w)^{-1} (\mathbf{X}(j) \\
&\quad - \mathbf{U}_k(j)) \Bigg\} + \log(P(C_k)) \\
&= \sum_{j=1}^{col} \Bigg\{ -\frac{1}{2} \log |\tilde{\mathbf{S}}_w| - \frac{1}{2}(\mathbf{X}(j))^{\mathrm{T}} (\tilde{\mathbf{S}}_w)^{-1} \mathbf{X}(j) \\
&\quad + (\mathbf{U}_k(j))^{\mathrm{T}} (\tilde{\mathbf{S}}_w)^{-1} \mathbf{X}(j) \\
&\quad - \frac{1}{2}(\mathbf{U}_k(j))^{\mathrm{T}} (\tilde{\mathbf{S}}_w)^{-1} \mathbf{U}_k(j) \Bigg\} + \log(P(C_k)).
\end{aligned}
$$

By eliminating the ineffective terms again, the Bayes classifier $g_k(\mathbf{X})$ could be further reduced and formulated as

$$g_k(\mathbf{X})= \sum_{j=1}^{col}\left\{(\mathbf{U}_k(j))^{\mathrm{T}}(\tilde{\mathbf{S}}_w)^{-1}\mathbf{X}(j)-\frac{1}{2}(\mathbf{U}_k(j))^{\mathrm{T}}(\tilde{\mathbf{S}}_w)^{-1}\mathbf{U}_k(j)\right\} + \log(P(C_k)) \tag{B.2}$$

Therefore, for two-class classification, it is said $\mathbf{X}\in C_1$ if and only if $g_1(\mathbf{X})>g_2(\mathbf{X})$, i.e.,

$$\sum_{j=1}^{col}\left\{(\mathbf{U}_1(j))^{\mathrm{T}}(\tilde{\mathbf{S}}_w)^{-1}\mathbf{X}(j)-\frac{1}{2}(\mathbf{U}_1(j))^{\mathrm{T}}(\tilde{\mathbf{S}}_w)^{-1}\mathbf{U}_1(j)\right\} + \log(P(C_1)) > \sum_{j=1}^{col}\left\{(\mathbf{U}_2(j))^{\mathrm{T}}(\tilde{\mathbf{S}}_w)^{-1}\mathbf{X}(j)-\frac{1}{2}(\mathbf{U}_2(j))^{\mathrm{T}}(\tilde{\mathbf{S}}_w)^{-1}\mathbf{U}_2(j)\right\} + \log(P(C_2)).$$

Then we could say $\mathbf{X}\in C_1$ if and only if

$$\sum_{j=1}^{col}\mathbf{w}_j^{\mathrm{T}}\mathbf{X}(j)+\mathbf{w}_0>0,$$

$$\mathbf{w}_j=(\tilde{\mathbf{S}}_w)^{-1}(\mathbf{U}_1(j)-\mathbf{U}_2(j)),$$

$$\mathbf{w}_0=\sum_{j=1}^{col}-\frac{1}{2}(\mathbf{U}_1(j)+\mathbf{U}_2(j))^{\mathrm{T}}(\tilde{\mathbf{S}}_w)^{-1}(\mathbf{U}_1(j)-\mathbf{U}_2(j))+\log\frac{P(C_1)}{P(C_2)}. \tag{B.3}$$

Finally, under the condition (3) in the theorem, i.e., $\Delta\mathbf{U}=s_i(\mathbf{U}_1(i)-\mathbf{U}_2(i))=s_j(\mathbf{U}_1(j)-\mathbf{U}_2(j)), \forall i\neq j, i, j=1,\ldots,col$, we then obtain the declaration that $\mathbf{X}\in C_1$ if and only if

$$\mathbf{w}_{bayes}^{\mathrm{T}}\left(\sum_{j=1}^{col}(s_j)^{-1}\mathbf{X}(j)\right)+\mathbf{w}_0>0,$$

$$\mathbf{w}_{bayes}=\mathbf{w}_1=\cdots=\mathbf{w}_{col}=(\tilde{\mathbf{S}}_w)^{-1}\Delta\mathbf{U} \tag{B.4}$$

else $\mathbf{X}\in C_2$.

Next, the following shows why 2D-LDA in terms of equality (3) would be a Bayes optimal feature extractor for two-class classification problem under the conditions indicated in Theorem 2. First, for two-class classification problem, $\mathbf{S}_b^{2d}=\sum_{j=1}^{col}\mathbf{S}_{b,j}^{2d}$ and $\mathbf{S}_w^{2d}=\sum_{j=1}^{col}\mathbf{S}_{w,j}^{2d}$, where

$$\mathbf{S}_{b,j}^{2d}=\frac{N_1}{N}(\mathbf{U}_1(j)-\mathbf{U}(j))(\mathbf{U}_1(j)-\mathbf{U}(j))^{\mathrm{T}}+\frac{N_2}{N}(\mathbf{U}_2(j)-\mathbf{U}(j))(\mathbf{U}_2(j)-\mathbf{U}(j))^{\mathrm{T}},$$

$$\mathbf{U}(j)=\frac{N_1}{N}\mathbf{U}_1(j)+\frac{N_2}{N}\mathbf{U}_2(j),$$

$$\mathbf{S}_{w,j}^{2d}=\frac{1}{N}\sum_{k=1}^{2}\sum_{i=1}^{N_k}(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))(\mathbf{X}_i^k(j)-\mathbf{U}_k(j))^{\mathrm{T}},\quad j=1,\ldots,col.$$

Note that $\mathbf{S}_{b,j}^{2d}$ and $\mathbf{S}_b^{2d}$ could be written equivalently below based on $N=N_1+N_2$ and equality (25):

$$\mathbf{S}_{b,j}^{2d}=\frac{N_1N_2}{N^2}(\mathbf{U}_1(j)-\mathbf{U}_2(j))(\mathbf{U}_1(j)-\mathbf{U}_2(j))^{\mathrm{T}}=\frac{N_1N_2}{N^2}(s_j)^{-2}\Delta\mathbf{U}(\Delta\mathbf{U}^{\mathrm{T}}),$$

$$\mathbf{S}_b^{2d}=\sum_{j=1}^{col}\mathbf{S}_{b,j}^{2d}=\frac{N_1N_2}{N^2}\Delta\mathbf{U}\left(\sum_{j=1}^{col}(s_j)^{-2}\Delta\mathbf{U}^{\mathrm{T}}\right).$$

Second, it is known that the optimal feature of 2D-LDA in terms of equality (3) for two-class classification problem would satisfy $\lambda_{opt}\mathbf{w}_{opt}^{2d}=(\mathbf{S}_w^{2d})^{-1}\mathbf{S}_b^{2d}\mathbf{w}_{opt}^{2d}$, $\lambda_{opt}>0$. Hence we have

$$\mathbf{w}_{opt}^{2d}=(\lambda_{opt})^{-1}(\mathbf{S}_w^{2d})^{-1}\frac{N_1N_2}{N^2}\Delta\mathbf{U}\left(\sum_{j=1}^{col}(s_j)^{-2}\Delta\mathbf{U}^{\mathrm{T}}\right)\mathbf{w}_{opt}^{2d}. \tag{B.5}$$

Since $(\lambda_{opt})^{-1}\frac{N_1N_2}{N^2}(\sum_{j=1}^{col}(s_j)^{-2}\Delta\mathbf{U}^{\mathrm{T}})\mathbf{w}_{opt}^{2d}$ is a scalar value, then we have

$$\mathbf{w}_{opt}^{2d}\propto(\mathbf{S}_w^{2d})^{-1}\Delta\mathbf{U}. \tag{B.6}$$

Furthermore, it is easy to verify $\mathbf{S}_w^{2d}=col\cdot\tilde{\mathbf{S}}_w$. Comparing with equality (B.4), we then have

$$\mathbf{w}_{opt}^{2d}\propto\mathbf{w}_{bayes}. \tag{B.7}$$

It means the discriminant feature of 2D-LDA is in proportion to the Bayes optimal feature obtained in equality (B.4). They are the same except some scalar scaling under the conditions indicated by the theorem.

References

[1] M. Kirby, L. Sirovich, Application of the KL procedure for the characterization of human faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (1990) 103–108.

[2] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991) 71–86.

[3] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 228–233.

[4] K. Fukunnaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, New York, 1991.

[5] A.R. Webb, Statistical Pattern Recognition, second ed., Wiley, New York, 2002.

[6] R.A. Fisher, The use of multiple measures in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.

[7] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.

[8] P.N. Belhumeur, J. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[9] P. Comon, Independent component analysis, a new concept?, Signal Processing 36 (1994) 287–314.

[10] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[11] P.C. Yuen, J.H. Lai, Face representation using independent component analysis, Pattern Recognition 35 (6) (2002) 1247–1257.

[12] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face recognition by independent component analysis, IEEE Trans. Neural Networks 13 (6) (2002) 1450–1464.

[13] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[14] S.Z. Li, X.W. Hou, H.J. Zhang, Learning spatially localized, parts-based representation, in: CVPR, 2001.
[15] A. Pascual-Montano, J.M. Carazo, K. Kochi, D. Lehmann, R.D. Pascual-Marqui, Nonsmooth nonnegative matrix factorization (nsNMF), IEEE Trans. Pattern Anal. Mach. Intell. 28 (3) (2006) 403–415.
[16] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[17] B. Moghaddam, Principle manifolds and probabilistic subspace for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 24 (6) (2002) 780–788.
[18] B. Moghaddam, T. Jebara, A. Pentland, Bayesian face recognition, Pattern Recognition 33 (2000) 1771–1782.
[19] J. Yang, D. Zhang, A.F. Frangi, J. Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 26 (1) (2004) 131–137.
[20] J. Yang, J.Y. Yang, From image vector to matrix: a straightforward image projection technique—IMPCA vs. PCA, Pattern Recognition 35 (9) (2002) 1997–1999.
[21] M. Li, B. Yuan, 2D-LDA: a novel statistical linear discriminant analysis for image matrix, Pattern Recognition Lett. 26 (5) (2005) 527–532.
[22] H. Xiong, M.N.S. Swamy, M.O. Ahmad, Two-dimensional FLD for face recognition, Pattern Recognition 38 (2005) 1121–1124.
[23] J. Ye, R. Janardan, Q. Li, Two-dimensional linear discriminant analysis, in: NIPS, 2004.
[24] H. Kong, L. Wang, E.K. Teoh, J.G. Wang, V. Ronda, Generalized 2D principal component analysis, in: IEEE Conference on IJCNN, Canada, 2005.
[25] J. Ye, Generalized low rank approximations of matrices, Mach. Learn. 61 (2005) 167–191.
[26] D. Zhang, Z.-H. Zhou, S. Chen, Diagonal principal component analysis for face recognition, Pattern Recognition 39 (2006) 140–142.
[27] J. Yang, D. Zhang, X. Yong, J.-Y. Yang, Two-dimensional discriminant transform for face recognition, Pattern Recognition 38 (2005) 1125–1129.
[28] H. Kong, L. Wang, E.K. Teoh, A framework of 2D Fisher discriminant analysis: application to face recognition with small number of training samples, in: CVPR, 2005.
[29] S. Noushatha, G. Hemantha Kumar, P. Shivakumara, $(2D)^2$ LDA: an efficient approach for face recognition, Pattern Recognition 39 (2006) 1396–1400.
[30] X.-Y. Jing, H.-S. Wong, D. Zhang, Face recognition based on 2D Fisherface approach, Pattern Recognition, 39 (2006) 707–710.
[31] N.V. Chawla, K. Bowyer, Random subspaces and subsampling for 2-D face recognition, in: CVPR, 2005.
[32] K. Liu, Y.-Q. Cheng, J.-Y. Yang, Algebraic feature extraction for image recognition based on an optimal discriminant criterion, Pattern Recognition 26 (6) (1993) 903–911.
[33] L. Wang, X. Wang, J. Feng, On image matrix based feature extraction algorithms, IEEE Trans. Syst Man Cybern.—Part B: Cybern. 36 (1) (2006) 194–197.
[34] D. Xu, S. Yan, L. Zhang, M. Li, W. Ma, Z. Liu, H. Zhang, Parallel image matrix compression for face recognition, in: Proceedings of the 11th International Multimedia Modelling Conference, 2005.
[35] L. Chen, H. Liao, M. Ko, J. Lin, G. Yu, A new LDA based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.
[36] R. Huang, Q.S. Liu, H.Q. Lu, S.D. Ma, Solving the small sample size problem of LDA, in: ICPR, 2002.
[37] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, Discriminative common vectors for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005) 4–13.
[38] J. Ye, R. Janardan, C.H. Park, H. Park, An optimization criterion for generalized discriminant analysis on undersampled problems, IEEE Trans. Pattern Anal. Mach. Intell. 26 (8) (2004) 982–994.
[39] M. Loog, R.P.W. Duin, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, IEEE Trans. Pattern Anal. Mach. Intell. 26 (6) (2004) 732–739.
[40] W.-S. Zheng, J.-H. Lai, P.C. Yuen, GA-fisher: a new LDA-based face recognition algorithm with selection of principal components, IEEE Trans. Syst. Man Cybern. Part B 35 (5) (2005) 1065–1078.
[41] C. Liu, H. Wechsler, Enhanced Fisher linear discriminant models for face recognition, in: ICPR, 1998.
[42] H. Yu, J. Yang, A direct LDA algorithm for high dimensional data with application to face recognition, Pattern Recognition 34 (10) (2001) 2067–2070.
[43] J. Ye, Q. Li, LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation, Pattern Recognition 37 (2004) 851–854.
[44] W. Zhao, R. Chellappa, P.J. Phillips, Subspace linear discriminant analysis for face recognition, Technical Report CAR-TR-914, CS-TR-4009, University of Maryland, College Park, MD.
[45] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition, Pattern Recognition Letters 26 (2005) 181–191.
[46] D.-Q. Dai, P.C. Yuen, Regularized discriminant analysis and its application to face recognition, Pattern Recognition 36 (2003) 845–847.
[47] P. Zhang, J. Peng, N. Riedel, Discriminant analysis: a least squares approximation view, in: CVPR, 2005.
[48] Z. Jin, J.Y. Yang, Z.S. Hu, Z. Lou, Face recognition based on the uncorrelated discriminant transformation, Pattern Recognition 34 (2001) 1405–1416.
[49] J. Duchene, S. Leclercq, An optimal transformation for discriminant and principal component analysis, IEEE Trans. Pattern Anal. Mach. Intell. 10 (6) (1988) 978–983.
[50] J.R. Price, T.F. Gee, Face recognition using direct, weighted linear discriminant analysis and modular subspaces, Pattern Recognition 38 (2005) 209–219.
[51] H.-C. Kim, D. Kim, S.Y. Bang, Face recognition using LDA mixture model, Pattern Recognition Lett. 24 (2003) 2815–2821.
[52] T.-K. Kim, J. Kittler, Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 318–327.
[53] F. De la Torre Frade, T. Kanade, Multimodal oriented discriminant analysis, in: International Conference on Machine Learning (ICML), August, 2005.
[54] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, H.-J. Zhang, Discriminant analysis with tensor representation, in: CVPR, 2005.
[55] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The feret evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[56] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1618.
[57] A.M. Martínez, M. Zhu, Where are linear feature extraction methods applicable?, IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1934–1944.
[58] J. Yang, J.Y. Yang, Why can LDA be performed in PCA transformed space?, Pattern Recognition 36 (2) (2003) 563–566.

About the Author—WEI-SHI ZHENG was born in Guangzhou (Canton), China, in 1981. He received the B.S. degree in both mathematics and computer science at Sun Yat-sen University, Guangzhou, in 2003. He is now pursuing a Ph.D. degree in applied mathematics at Sun Yat-Sen University. From April 2006 to October 2006, he was a visiting student at Center for Biometrics and Security Research and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. In 2007, he has been an exchanged research student at Hong Kong Baptist University from May 16 to November 15. He is a student member of IEEE.
His current research interests include pattern recognition, machine learning, and face recognition.

About the Author—J.H. LAI was born in 1964. He received the M.Sc. degree in applied mathematics in 1989 and the Ph.D. degree in mathematics in 1999 from Sun Yat-sen University, Guangzhou, China. He joined Sun Yat-sen University in 1989, where currently, he is a Professor with the Department of Electronics and Communication Engineering, School of Information Science and Technology. He has published over 50 papers in the international journals, book chapters, and conferences. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelets and their applications. Dr. Lai had successfully organized the International Conference on Advances in Biometric Personal Authentication' 2004, which was also the Fifth Chinese Conference on Biometric Recognition (Sinobiometrics'04), Guangzhou, in December 2004. He has taken charge of more than four research projects, including NSFC (number 60144001, 60 373 082, 60675016), the Key (Key grant) Project of Chinese Ministry of Education (number 105 134), and NSF of Guangdong, China (number 021 766, 06023194). He serves as a board member of the Image and Graphics Association of China and also serves as a board member and secretary-general of the Image and Graphics Association of Guangdong.

About the Author—STAN Z. LI received the Ph.D. degree from Surrey University, UK, in 1991. He is currently a professor at National Laboratory of Pattern Recognition (NLPR), director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA); and director of Joint Laboratory for Intelligent Surveillance and Identification in Civil Aviation (CASIA-CAUC). He worked at Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor at Nanyang Technological University, Singapore. His research interest includes face recognition, biometrics, intelligent video surveillance, pattern recognition and machine learning, and image and video processing. He authored the book "Markov Random Field Modeling in Image Analysis" (Springer, 1st edition in 1995 and 2nd edition in 2001), co-edited "Handbook of Face Recognition" (Springer, 2005), and published over 200 papers in international journals and conferences. He is currently an associated editor of IEEE Transactions on Pattern Analysis and Machine Intelligence. He leads several national and international collaboration projects in biometrics and intelligent video surveillance. He has been actively participating in organizing a number of international conferences and workshops in the fields of computer vision, image processing, pattern recognition, face analysis and biometrics.