Pattern Recognition Letters 29 (2008) 2122–2128

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A note on two-dimensional linear discriminant analysis

Zhizheng Liang [a,*], Youfu Li [a], Pengfei Shi [b]

[a] Manufacturing Engineering and Engineering Management, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
[b] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, China

ARTICLE INFO

Article history:
Received 22 August 2007
Received in revised form 22 June 2008
Available online 5 August 2008

Communicated by F. Roli

Keywords:
Feature extraction
Linear discriminant analysis
2DLDA
Discriminant power
Distance measure

ABSTRACT

2DLDA and its variants have attracted much attention from researchers recently due to the advantages over the singularity problem and the computational cost. In this paper, we further analyze the 2DLDA method and derive the upper bound of its criterion. Based on this upper bound, we show that the discriminant power of two-dimensional discriminant analysis is not stronger than that of LDA under the assumption that the same dimensionality is considered. In experimental parts, on one hand, we confirm the validity of our claim and show the matrix-based methods are not always better than vector-based methods in the small sample size problem; on the other hand, we compare several distance measures when the feature matrices and feature vectors are applied. The matlab codes used in this paper are available at http://www.mathworks.com/matlabcentral/fileexchange/loadCategory.do?objectType=category&objectId=127&objectName=Application.

1. Introduction

Feature extraction plays an important role in pattern recognition and computer vision, which has been widely used for face recognition and character recognition. In general, the aim of feature extraction is to reduce the dimensionality of data so that the extracted features are as representable as possible. During the past several decades, a number of algorithms for feature extraction have been developed, such as principal component analysis (PCA) and linear discriminant analysis (LDA). PCA is to find an orthogonal set of vectors that maximize the variance of the projected vectors, while LDA is to seek the discriminant vectors such that the ratio of the between-class distance to the within-class distance is maximized. However, LDA often suffers from the small sample size (3S) problem when the dimension of data is much bigger than the number of data points. To deal with this problem, some effective approaches have been proposed, including regularized LDA, PCA + LDA (Belhumeur et al., 1997), pseudo-inverse LDA, orthogonal LDA (Ye, 2005a), LDA/GSVD (Howland and Park, 2004), and LDA/QR (Ye and Li, 2005b). The PCA + LDA method (Belhumeur et al., 1997) is one of the most popular methods, whose idea is to apply PCA to reduce the dimension of data before performing LDA. Further, Ye et al. (2004a,b) proposed generalized discriminant analysis in terms of a new optimization criterion which can overcome the singularity problem. Based on this criterion, orthogonal LDA (Ye, 2005a) whose discriminant vectors are orthogonal is developed. It is shown that orthogonal LDA is superior to uncorrelated LDA in some cases. The LDA/GSVD algorithm (Howland and Park, 2004) circumvents the singularity problem by using the generalized singular value decomposition. The LDA/QR algorithm (Ye and Li, 2005b) first applies the QR decomposition on a small matrix, and then followed by LDA. All the algorithms contribute to the development of LDA.

In recent years, there has interest in developing matrix-based methods for data representation and classification. The essence of these methods is that the image data is not converted into vectors prior to dimensionality reduction. Two-dimensional PCA (Yang et al., 2004b) is an effective image presentation and recognition technique. It first constructs the image covariance matrix and then performs the eigen-decomposition on this covariance matrix. The GLRAM algorithm (Ye, 2005c) generalizes 2DPCA in some sense, which stores fewer coefficients for images than 2DPCA. Based on similar ideas in 2DPCA, one-sided 2DLDA is developed for the classification tasks. For example, Li and Yuan (2005) proposed 2DLDA which directly extracts the features from the images based on Fisher's criterion. Similar ideas can be found in Kongsontana and Rangsanseri (2005), Xiong et al. (2005). Uncorrelated image matrix-based linear discriminant analysis (MLDA) (Yang et al., 2003a) is a variant of 2DLDA. MLDA eliminates the correlation between discriminant vectors and obtains better recognition performance than LDA in some cases. Cho et al. (2006) noted that one-sided 2DLDA essentially works by the row-direction of images. Based on this, they proposed two-direction 2DLDA by considering row and column directions. Two-direction LDA stores fewer

* Corresponding author. Fax: +852 2788 8423.
E-mail addresses: zhliang@cityu.edu.hk (Z. Liang), meyfli@cityu.edu.hk (Y. Li), pfshi@sjtu.edu.cn (P. Shi).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.07.009

coefficients than one-sided LDA. Other related work can be found in Kong et al. (2005), Noushath et al. (2006), Yang et al. (2005c). Different from the above ideas, Ye et al. (2004a,b) developed another variant of 2DLDA. Their method works by an iterative way. All the algorithms can overcome the singularity problem in the small sample size problem and achieve higher computational efficiency than classical LDA. In addition, Wang et al. (2006) found that two-dimensional 2DLDA is a special block-based method such as column-based or row-based methods in essence. Overall, all the algorithms contribute to the development of 2DLDA.

Although 2DLDA and its some variants have advantages over the singularity problem and the computational cost, we theoretically prove that the discriminant power of 2DLDA is not stronger than that of LDA when considering the same dimensionality in this paper. Moreover, we show by experiments on handwritten numerical characters that the recognition performance of 2DLDA is not always better than that of LDA when the rank of the within-class scatter matrix approaches the dimension of data. In addition, we compare several distance measures when matrix-based methods and vector-based methods are adopted on face databases and show the matrix-based methods are not always better than vector-based methods in the small sample size problem.

2. Related work

2.1. Linear discriminant analysis

Let us consider a set of m training samples $\{x_1, x_2, \ldots, x_m\}$ taking values in an n-dimensional space. Let L be the number of classes and l_i be the number of training samples of class i, where $i = 1, \ldots, L$. Then the between-class scatter matrix S_b and the within-class scatter matrix S_w are constructed as follows:

$$S_b = \sum_{i=1}^{L} l_i (m_i - m_0)(m_i - m_0)^T, \tag{1}$$

$$S_w = \sum_{i=1}^{L} \sum_{j=1}^{l_i} (x_i^{(j)} - m_i)(x_i^{(j)} - m_i)^T, \tag{2}$$

where $x_i^{(j)}$ is the jth sample of class i, m_i $(i = 1, \ldots, L)$ is the mean vector of training samples in class i, and m_0 is the mean vector of all training samples.

The LDA method tries to find the projected matrix that maximizes the ratio of the between-class distance to the within-class distance in the projected space:

$$J_1(W) = \max \frac{\operatorname{trace}(W^T S_b W)}{\operatorname{trace}(W^T S_w W)}, \tag{3}$$

where W is an $n \times q$ matrix whose columns consist of q discriminant vectors.

Eq. (3) has explicit semantics for both numerator and denominator. That is, $\operatorname{trace}(W^T S_b W)$ measures the separation between classes in the projected space and $\operatorname{trace}(W^T S_w W)$ measures the closeness of the vectors within the classes in the projected space.

Instead of adopting Eq. (3), some researchers often search the most discriminant vectors by the following classical LDA criterion:

$$J_2(W) = \max \operatorname{trace}((W^T S_w W)^{-1} W^T S_b W). \tag{4}$$

It is not difficult to verify that Eq. (4) is invariant under any nonsingular linear transformation. Unlike Eqs. (3) and (4) does not have explicit semantics. In addition to Eqs. (3) and (4), there are other variants of LDA (Fukunaga, 1990). Applying a joint diagonalization (Fukunaga, 1990), we can solve Eq. (4). That is, finding the discriminant matrix W is to diagonalize simultaneously both matrices S_b and S_w (i.e. $W^T S_w W = I$, $W^T S_b W = \Lambda$, where I is an identity matrix and Λ is a diagonal matrix whose elements are in decreasing order). Directly solving Eq. (3) is intractable. Often Eq. (3) is approximately solved by the generalized eigenvalue problem $S_b w_i = \lambda_i S_w w_i$, where is the eigenvector corresponding to the ith largest eigenvalue λ_i and w_i constitutes the ith column of the matrix W.

2.2. Two-dimensional linear discriminant analysis

Let $\{A_1, \ldots, A_m\}$ be m image matrices, where A_i $(i = 1, \ldots, L)$ are $r \times c$ matrices. Let M_i $(i = 1, \ldots, L)$ be the mean image of training samples in class i and M_0 be the mean image of all training samples. Consider a $s \times t$-dimensional space $L \otimes R$, where $\otimes$ denotes the tensor product, L is spanned by $\{u_1, \ldots, u_s\}$, and R is spanned by $\{v_1, \ldots, v_t\}$. Let us define two matrices: $L = [u_1 \cdots u_s]$ and $R = [v_1 \cdots v_t]$.

Considering the case of image data, the feature extraction method is to find L and R such that the original image space A_i is converted into a low-dimensional image space by $B_i = L^T A_i R$. In the low-dimensional space obtained by the linear transformations L and R, the between-class distance D_b and the within-class distance D_w are defined as follows:

$$D_b = \sum_{i=1}^{L} l_i \left\| L^T (M_i - M_0) R \right\|_F^2, \tag{5}$$

$$D_w = \sum_{i=1}^{L} \sum_{j=1}^{l_i} \left\| L^T (A_i^{(j)} - M_i) R \right\|_F^2, \tag{6}$$

where $A_i^{(j)}$ denotes the jth training sample in class i, and $\| \|_F$ denotes the Frobenius norm.

Observe that $\|A\|_F^2 = P\operatorname{trace}(A^T A) = \operatorname{trace}(AA^T)$ for any matrix A. In such a case, Eqs. (5) and (6) can further be represented as

$$D_b = \operatorname{trace}\left(\sum_{i=1}^{L} l_i L^T (M_i - M_0) R R^T (M_i - M_0)^T L \right), \tag{7}$$

$$D_w = \operatorname{trace}\left(\sum_{i=1}^{L} \sum_{j=1}^{l_i} L^T (A_i^{(j)} - M_0) R R^T (A_i^{(j)} - M_0)^T L \right). \tag{8}$$

Similar to LDA, the 2DLDA method is to find matrices L and R such that the class structure of the original space is persevered in the projected space. So the criterion can be defined as

$$J_3(L, R) = \max \frac{D_b}{D_w}. \tag{9}$$

It is obvious that Eq. (9) contains the transformation matrices L and R. The optimal transformation matrices L and R can be obtained by maximizing D_b and minimizing D_w. However, it is very difficult to compute the optimal L and R simultaneously. In Ye's paper (Ye et al., 2004a,b), two optimization functions are defined to obtain L and R. For a fixed R, L is obtained by solving the following optimization function:

$$J_4(L) = \max \operatorname{trace}((L^T S_w^R L)^{-1} (L^T S_b^R L)) \tag{10}$$

where

$$S_b^R = \sum_{i=1}^{L} l_i (M_i - M_0) R R^T (M_i - M_0)^T \text{ and}$$

$$S_w^R = \sum_{i=1}^{L} \sum_{j=1}^{l_i} (A_i^{(j)} - M_0) R R^T (A_i^{(j)} - M_0)^T.$$

For a fixed L, R is obtained by solving the following optimization function:

$$J_5(R) = \max \operatorname{trace}((R^T S_w^L R)^{-1} (R^T S_b^L R)) \tag{11}$$

where $S_b^L = \sum_{i=1}^{L} l_i (M_i - M_0)^T L L^T (M_i - M_0)$ and

$$S_w^L = \sum_{i=1}^{L}\sum_{j=1}^{l_i}(A_i^{(j)} - M_0)^T L L^T (A_i^{(j)} - M_0).$$

To be specific, for a fixed R, the optimal L can be obtained by solving a generalized eigenvalue problem from Eq. (10). Similarly, R can be obtained by solving a generalized eigenvalue problem from Eq. (11) in the case of a fixed L. These two steps are repeated until the defined criterion is met.

3. The theoretical analysis between 2DLDA and LDA

In this section, we analyze 2DLDA and show the relationship among Eqs. (3), (4), (9)–(11). Furthermore, we demonstrate that the discriminant power of two-dimensional linear discriminant analysis is not stronger than that of LDA under the assumption that the same dimensionality is considered. In addition, we also point out the relationship between 2DLDA and LDA.

In the following, we start by stating the following lemma. For completeness, we also give its proof.

Lemma 1. (Wang et al., 2003). *Let B be a positive definite Hermite matrix and A be a positive semidefinite Hermite matrix. Then one has*

$$\frac{\mathrm{tr}(A)}{\mathrm{tr}(B)} \leqslant \mathrm{tr}\left(\frac{A}{B}\right). \tag{12}$$

Proof. Let $\lambda_1(B)$ denote the maximal eigenvalue of B, $A^{\frac{1}{2}}$ be the square root of A, and I be an identity matrix. Because $A^{\frac{1}{2}}(\lambda_1(B)I - B)A^{\frac{1}{2}} \geqslant 0$, one has $A^{\frac{1}{2}}\lambda_1(B)A^{\frac{1}{2}} \geqslant A^{\frac{1}{2}}BA^{\frac{1}{2}}$. Further, $\lambda_1(B)A \geqslant A^{\frac{1}{2}}BA^{\frac{1}{2}}$. Note that $\mathrm{tr}(BA) = \mathrm{tr}(BA)$. Then $\mathrm{tr}(\lambda_1(B)A) \geqslant \mathrm{trace}(BA)$. From this derivation, one has $\mathrm{tr}(A) = \mathrm{tr}(BB^{-1}A) \leqslant \mathrm{tr}(\lambda_1(B)B^{-1}A) = \lambda_1(B)\mathrm{tr}(B^{-1}A) \leqslant \mathrm{tr}(B)\mathrm{tr}(B^{-1}A)$.

It follows that $\mathrm{tr}(A) \leqslant \mathrm{tr}(B)\mathrm{tr}(B^{-1}A)$.

This completes the proof. □

From Lemma 1, it is not difficult to obtain the following propositions.

Proposition 1. *Assume $J_1(W)$ and $J_2(W)$ are given in terms of Eqs. (3) and (4). Then one has*

$$J_1(W) \leqslant J_2(W). \tag{13}$$

Proof. It is a straightforward matter to show that Eq. (13) holds from Eq. (12). □

Proposition 2. *Assume $J_3(L,R)$ and $J_4(L)$ are given in terms of Eqs. (9) and (10). Then one has*

$$J_3(L,R) \leqslant J_4(L). \tag{14}$$

Proposition 3. *Assume $J_3(L,R)$ and $J_5(L)$ are given in terms of Eqs. (9) and (11). Then one has*

$$J_3(L,R) \leqslant J_5(L). \tag{15}$$

We are now ready to prove the following theorem that will play a key role in this note.

Theorem 1. *Assume $J_1(W)$ and $J_3(L,R)$ are given in terms of Eqs. (3) and (9) and are used to extract features from image data. If $q = s \times t$, then $J_3(L,R) \leqslant J_1(W)$.*

Proof. From Eq. (5), one has

$$D_b = \sum_{i=1}^{L} l_i \left\| L^T(M_i - M_0)R \right\|_F^2 = \sum_{i=1}^{L} l_i \left\| \mathrm{vec}(L^T(M_i - M_0)R) \right\|_2^2, \tag{16}$$

where vec() denotes the vec operator which can convert the matrix into a vector by stacking the columns of the matrix.

For matrices A, B, and C, one has

$$\mathrm{vec}(ABC) = (C^T \otimes A)\mathrm{vec}(B). \tag{17}$$

For any column vector x, one has

$$\|x\|_2^2 = x^T x = \mathrm{trace}(xx^T). \tag{18}$$

Applying Eqs. (17) and (18), one can rewrite Eq. (16) as follows:

$$\sum_{i=1}^{L} l_i \|\mathrm{vec}(L^T(M_i - M_0)R)\|_2^2 = \sum_{i=1}^{L} l_i \|(R^T \otimes L^T)\mathrm{vec}(M_i - M_0)\|_2^2. \tag{19}$$

For the sake of notational simplicity, let $W^T = (R^T \otimes L^T)$. It is straightforward to verify that W is an $rc \times st$ matrix. Then Eq. (19) is further represented as

$$\begin{aligned}\sum_{i=1}^{L} l_i \left\| (R^T \otimes L^T)\mathrm{vec}(M_i - M_0) \right\|_2^2 &= \mathrm{trace}\Big(W^T \sum_{i=1}^{L} l_i (m_i - m_0) \\ &\quad \times (m_i - m_0)^T W\Big) \\ &= \mathrm{trace}(W^T S_b W).\end{aligned} \tag{20}$$

It is clear that Eq. (20) has the form of the numerator of the fraction in Eq. (3). Likewise, from Eq. (6), one has

$$\begin{aligned} D_w &= \sum_{i=1}^{L}\sum_{j=1}^{l_i} \left\| L^T(A_i^{(j)} - M_i)R \right\|_F^2 \\ &= \sum_{i=1}^{L}\sum_{j=1}^{l_i} \left\| (R^T \otimes L^T)\mathrm{vec}(A_i^{(j)} - M_i) \right\|_2^2 \\ &= \mathrm{trace}\Big(W^T \sum_{i=1}^{L}\sum_{j\in 1}^{l_i} (x_i^{(j)} - m_i)(x_i^{(j)} - m_i)^T W\Big) \\ &= \mathrm{trace}(W^T S_w W). \end{aligned} \tag{21}$$

It is clear that Eq. (21) has the form of the denominator of the fraction in Eq. (3).

From Eqs. (20) and (21), one has

$$\frac{D_b}{D_w} = \frac{\mathrm{trace}(W^T S_b W)}{\mathrm{trace}(W^T S_w W)}. \tag{22}$$

Note that W in Eq. (22) is an $rc \times st$ matrix and W in Eq. (3) is an $n \times q$ matrix. If Eq. (3) is used to deal with image data, it is necessary to convert image data into vector data. In such a case, $n = r \times c$. Furthermore, when $q = s \times t$, we find that Eq. (22) has the form of Eq. (3) under the assumption that $W^T = (R^T \otimes L^T)$. As a result, the maximal value of $\frac{D_b}{D_w}$ is equal to $J_1(W)$ when $q = s \times t$ and $W^T = (R^T \otimes L^T)$, where W is the solution of Eq. (3), R and L are the solution of Eq. (9). However, it is observed that R and L are obtained by the iterative algorithm. Theoretically, R and L are locally optimal. As a result, $J_3(L,R)$ obtains the locally maximal value in the general cases. Assume that R and L are obtained by solving Eq. (9). Let $W_0^T = (R^T \otimes L^T)$. If W_0 is the solution of Eq. (3), then $J_3(L,R) = J_1(W)$. If W_0 is not the solution of Eq. (3), then we have $J_3(L,R) < J_1(W)$ due to Eq. (22). Since R and L are locally optimal by solving Eq. (9) in the general cases, we hope to use another scheme to obtain R and L. That is, we first solve Eq. (3) to obtain W. If there exist R and L such that $W^T = (R^T \otimes L^T)$, then $J_3(L,R) = J_1(W)$. However, decomposing W into the form of the Kronecker product of two matrices R and L is still difficult, even impossible. Consequently, obtaining R and L by this scheme may be impossible. From the above analysis, we know that the condition $W^T = (R^T \otimes L^T)$ is necessary for showing that $J_3(L,R) = J_1(W)$. In the case of not considering the condition $W^T = (R^T \otimes L^T)$, we always have $J_3(L,R) \leqslant J_1(W)$.

This completes the proof of the theorem. □

From Proposition 1 and Theorem 1, it is not difficult to obtain the following proposition.

Proposition 4. *Assume $J_2(W)$ and $J_3(L,R)$ are given in terms of Eqs. (4) and (9) and are used to extract features from image data. If $q = s \times t$, then $J_3(L,R) \leqslant J_2(W)$.*

Theorem 1 demonstrates the upper bound of D_b/D_win Eq. (9). It is obvious that if the solution of Eq. (3) can be decomposed into the Kronecker product of two matrices R and L with approximate dimensions, the value of the objective function in Eq. (9) is equal to the value of Eq. (3). To our knowledge, it is easy to compute the Kronecker product of two matrices. It is, however, very difficult or impossible to decompose a large matrix into the form of the Kronecker product of two matrices. As a result, the upper bound of D_b/D_w may not be obtained when L and R are matrices. In other words, the condition $W^T = (R^T \otimes L^T)$ does not hold. However, note that we can obtain the upper bound of D_b/D_w if A_i $(i = 1,\ldots,s)$ are $r \times 1$ matrices, namely, vectors. It is clear that this is the form of LDA. Therefore, in some sense, LDA is a special case of 2DLDA when the objective function achieves the optimal value. In addition, when one of L and R is an identity matrix, the corresponding L or R can be obtained by solving Eq. (9). Note that in Xiong et al. (2005), under the condition that L is set to an identity matrix, the corresponding algorithm is also referred to as two-dimensional discriminant analysis. It is clear that Ye's method (Ye et al., 2004) is a generalization of some so-called one-sided 2DLDA methods (Xiong et al., 2005). It is not difficult to show that the theorem still holds in one-sided 2DLDA. If Eq. (3) or Eq. (9) is used to define the discriminant power of LDA or 2DLDA, we can see that the discriminant power of 2DLDA or its variants is not stronger than that of LDA under the condition that the reduced dimension is equal.

4. Experimental results

This section reports a set of experiments and performs a comparison of several methods for classification tasks. Two types of data are applied: handwritten numerical characters and face images. All the algorithms are programmed in Matlab Language.

4.1. Experiments on handwritten numerical characters

In this section, the experiments on handwritten numerical characters are used to verify the proposed claim. The handwritten numeral characters are obtained from the UCI data Repository (Blake et al., 1998). This data set consists of 5620 handwritten numeral characters. The original image of each character has the size of 32×32 pixels. In order to reduce the size of the image, we obtain the size of 16×16 pixels where each pixel is obtained from the average of the block of 2×2 pixels in the original images. In other words, the number of features of each character image is 256. From the discussion in Section 3, we can see that two important parameters s and t are involved. For simplicity, s and t are set to the common value d in the following experiments. We know that the number of features is $s \times t$. In addition, when the classification problem is involved, we adopt the nearest neighbor classifier due to its simplicity.

The first set of experiments is used to show the discriminant power of LDA and 2DLDA. The corresponding results are shown in Table 1. As can be seen from Table 1, the discriminant power of LDA is stronger than that of 2DLDA when considering the same dimensions. It shows that the performance of 2DLDA may not be superior to that of LDA, which verifies our claim in Section 3. In addition, we carry out experiments to evaluate 2DLDA and LDA in terms of the classification performance. In order to reduce variation, the experimental results we report in this paper are averaged over 10 randomizations. Fig. 1 shows the experimental results, where the x axis denotes the number of the training samples in each class, and the y axis denotes the recognition performance. As observed in Fig. 1, LDA is superior to 2DLDA in terms of the classification performance when considering the same dimensions with the increase of training samples.

4.2. Experiments on the ORL face database

The ORL database (ORL, 1992) contains images from 40 individuals, each providing 10 different images. For some subjects, the images are taken at different times. The facial expressions (open or closed eyes, smiling or non-smiling) and facial details (glasses or no glasses) also vary. The images are taken with a tolerance for some tilting and rotation of the face of up to 20°. All images are grayscale and normalized to a resolution of 112×96 pixels. For the computational efficiency, we resize the images to 56×46 pixels. In our experiments, we use the nearest-neighbor (NN) algorithm as the classifier. It is found that the distance measures in the NN classifier affect the classification performance, especially for measuring the features in matrix form (Zuo et al., 2006; Meng and Zhang, 2007). For the features expressed in matrix form, the assembled matrix distance (AMD) metric (Zuo et al., 2006) and the volume measure (Meng and Zhang, 2007) are proposed to measure the distance between two feature matrices. These two distance measures are described as follows.

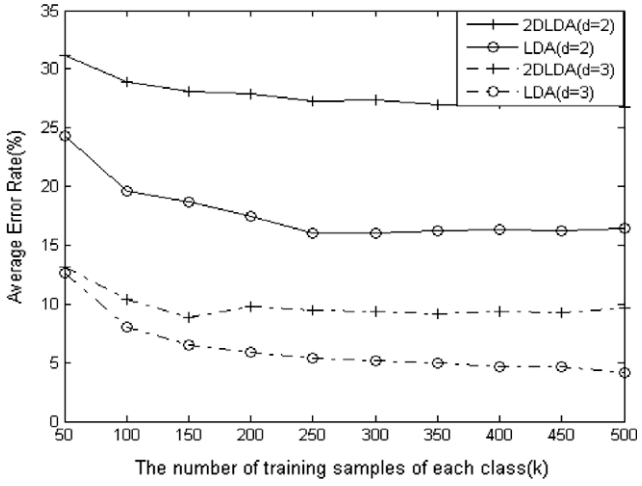

Fig. 1. Comparisons of performance with the change of training samples.

Table 1
Comparisons of discriminant power between LDA and 2DLDA

The training number	1000			3000			5000		
The dimensions	1	4	9	1	4	9	1	4	9
LDA	9.46	6.55	4.16	7.42	5.36	3.47	7.13	5.14	3.41
2DLDA	4.63	2.77	1.82	4.33	2.70	1.67	4.28	2.65	1.67

The assembled matrix distance is defined as $d_{\mathrm{AMD}}(C,D) = (\sum_{j=1}^{t}(\sum_{i=1}^{s}(a_{ij}-b_{ij})^2)^{(1/2)p})^{1/p}$, where $C=(c_{ij})_{s\times t}$ and $D=(d_{ij})_{s\times t}$ are two feature matrices and $p>0$.

In fact, the Frobenius distance measure is a special case of AMD metric with $p=2$.

The volume measure (VM) is defined as

$$\mathrm{vol}(C-D)=\sqrt{\det(C-D)^{\mathrm{T}}(C-D)},$$

where $C=(c_{ij})_{s\times t}$ and $D=(d_{ij})_{s\times t}$ are two feature matrices.

In contract, for two features in vector form, there exists the p-norm distance, defined as

$d_p=(\sum_{i=1}^{q}\|c_i-d_i\|^p)^{1/p}$ where $C=(c_i)_q$ and $D=(d_i)_q$ are two feature vectors.

It is obvious that d_p is the Euclidean distance measure with $p=2$. Since d_p and $d_{\mathrm{AMD}}(C,D)$ involve the parameter p. We search the optimal parameter in the interval of 0.1 and 3 by an increment of 0.1 and show the best result when $d_{\mathrm{AMD}}(C,D)$ and d_p are used. In addition, we list the results with the Frobenius distance measure and the Euclidean distance (ED) measure. For comparisons, we also implement some methods including the iterative 2DLDA method (Ye et al., 2004a,b), the one-sided 2DLDA method (Li and Yuan, 2005), the Fisherface method (Belhumeur et al., 1997) and the orthogonal LDA method (Ye, 2005a). In this set of experiments, a training sample set is formed by randomly selecting t (2–9) images from each individual and the remaining images are used for testing. To enhance the accuracy of performance, the classification performance reported in the experiments is averaged over twenty runs. In other words, twenty different training and testing sets are used for performance evaluation. In general, the performance of all methods we use varies with the number of dimensions. For comparisons, we explore the performance on all the feature dimensions and report the best results on all possible feature dimensions. Fig. 2 denotes the error rates of dimension reduction methods with the change of training sample numbers on the face database. For clarity, we also show the means and standard deviations in the parentheses of the error rates in Table 2.

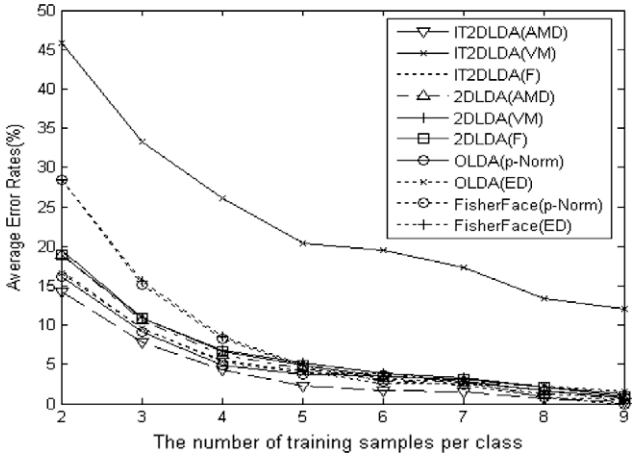

Fig. 2. Error rates of each method with the change of distance measures and training sample numbers.

From Fig. 2 and Table 2, we can see that the more the training samples we use, the smaller the error rate. The iterative 2DLDA method with the AMD measure or the Frobenius measure and the one-sided 2DLDA method are superior to the Fisherface method when the number of training samples per class is smaller than 5. In fact, in previous literature, 2DLDA is often compared with the Fisherface method and does not compare with some LDA methods which consider the null space of the within-class scatter matrix. From Table 3, it is found that the OLDA method is still superior to one-sided 2DLDA and is competitive with the iterative 2DLDA with the Frobenius measure in terms of the classification performance. A possible explanation is that the Fisherface method removes the null space of the within-class scatter matrix and the orthogonal LDA method does not remove the null space. The performance of the one-sided 2DLDA method with the volume measure is competitive with that of the one-sided 2DLDA method with other measures. The iterative 2DLDA method with the volume measure performs poorly. A possible explanation is that some columns of difference matrices obtained by dimension reduction are deleted to get full rank matrices in implementing VM classification and the full rank matrices lose some information of the reduced features. The one-sided 2DLDA method with the volume measure does not perform poorly because the full rank matrices still contain some important information of reduced features. It is also noted that 2DLDA with the AMD measure is superior to the corresponding 2DLDA method with the Frobenius measure since the AMD measure contains the Frobenius measure. Likewise, LDA with the p-norm measure is superior to LDA with the ED measure since the p-norm measure contains ED. The experiments show that choosing distance measures for feature matrices is important for the 2DLDA methods. Finally, we explain the reason that the iterative 2DLDA method achieves better than other methods when the number of training samples per class is small on this database. That is, although LDA has stronger discriminant power than 2DLDA, high Fisher's value does not definitely yield better generalization ability. Further speaking, if the within-class distance is very small and the between-class distance is also very small, then it may yield large Fisher values. But a smaller between-class distance may lead to the difficulty in recognizing different classes.

4.3. Experiments on the Yale face database

The last experiment is performed using the Yale face database (Yale, 1997), which contains 165 images of 15 individuals (each

Table 2
Comparisons of misclassification rates (%) for iterative 2DLDA, one-sided 2DLDA, Fisherface, and OLDA with the change of distance measures and training sample numbers on the ORL face database

Methods	Measures	2	3	4	5	6	7	8	9
Iterative 2DLDA	AMD	14.22 (2.71)	7.75 (2.85)	4.25 (2.09)	2.30 (1.09)	1.75 (1.17)	1.58 (1.00)	0.63 (0.8)	0.25 (0.79)
	Frobenius	16.84 (2.92)	9.54 (4.27)	5.46 (1.80)	4.15 (1.31)	2.62 (1.36)	2.58 (1.43)	1.25 (1.10)	0.75 (1.69)
	VM	45.61 (2.38)	33.2 (3.30)	26.0 (2.23)	20.4 (3.11)	19.5 (2.73)	17.3 (2.72)	13.3 (3.91)	12.0 (4.30)
One-sided 2DLDA	AMD	18.94 (3.07)	10.6 (2.89)	6.17 (1.87)	4.45 (1.23)	3.06 (1.60)	2.67 (1.29)	1.75 (1.58)	0.60 (1.05)
	Frobenius	18.94 (2.98)	10.8 (3.39)	6.71 (1.89)	4.85 (0.94)	3.44 (1.57)	3.17 (1.10)	2.13 (1.77)	1.00 (1.75)
	VM	19.53 (3.03)	10.8 (3.74)	6.79 (2.04)	5.25 (2.11)	3.88 (1.55)	3.25 (1.59)	2.13 (1.87)	0.50 (1.77)
OLDA	p-norm	16.19 (2.10)	9.18 (2.99)	4.88 (1.40)	3.70 (0.92)	3.56 (1.04)	2.83 (1.37)	1.75 (1.88)	0.75 (1.21)
	ED	16.59 (2.01)	9.67 (2.83)	5.37 (1.04)	3.90 (0.91)	3.87 (1.71)	2.91 (1.37)	2.25 (2.49)	0.50 (2.11)
Fisherface	p-norm	28.38 (3.01)	15.2 (2.84)	8.33 (1.85)	4.70 (1.40)	2.81 (1.77)	2.42 (0.92)	0.88 (1.03)	0.00 (0.00)
	ED	28.47 (2.03)	15.5 (2.81)	8.54 (1.66)	4.85 (1.91)	3.19 (2.19)	2.83 (1.19)	0.88 (1.03)	0.50 (1.05)

Table 3
Comparisons of misclassification rates (%) for iterative 2DLDA, one-sided 2DLDA, Fisherface, and OLDA with the change of distance measures and training sample numbers on the Yale face database

Methods	Measures	2	4	6	8	10
Iterative 2DLDA	AMD	55.37 (6.55)	32.90 (3.23)	25.73 (4.33)	20.66 (6.69)	16.33 (8.23)
	Frobenius	56.96 (5.11)	34.42 (5.85)	27.06 (4.96)	22.22 (7.88)	19.00 (13.1)
	VM	76.77 (5.60)	63.90 (3.02)	56.53 (4.66)	51.00 (7.03)	46.00 (10.8)
One-sided 2DLDA	AMD	50.55 (4.95)	31.90 (3.77)	25.26 (4.67)	21.11(5.66)	14.33 (7.26)
	Frobenius	52.20 (4.98)	33.74 (3.94)	26.67 (4.37)	23.12 (6.17)	19.00 (7.26)
	VM	50.80 (5.21)	34.76 (3.83)	25.31 (5.03)	24.78 (6.38)	19.33 (7.46)
OLDA	p-Norm	35.40 (5.14)	19.33 (4.15)	13.73 (2.97)	8.55 (3.40)	5.90 (5.24)
	ED	37.00 (5.40)	20.52 (4.25)	14.93 (3.58)	9.55 (3.68)	6.64 (7.17)
Fisherface	p-Norm	59.37 (5.32)	37.00 (2.73)	26.00 (4.46)	16.11 (4.61)	7.66 (5.83)
	ED	61.51 (6.93)	38.57 (2.61)	28.20 (4.82)	18.89 (5.12)	11.33 (8.94)

person has 11 different images) under various facial expressions and lighting conditions. Each image is manually cropped and resized to 56×46 pixels in this experiment. We consider this database in order to evaluate the performance of methods under the condition when facial expression and lighting conditions are changed.

For the Yale face database, we carry out the experiments which are similar to the ones in Section 4.2. We use the first t ($t = 2,4,6,8,10$) image samples per class for training and the remaining images for testing. Table 3 shows the detailed experimental results of each method in terms of distance measures.

As can be seen from Table 3, the error rates of each method with different measures decease with the increase of the training samples. The OLDA method achieves better classification results than other methods on this database. When the number of training samples is small, the Fisherface method is worse than 2DLDA with the AMD measure and the Frobenius measure, which is consistent with the results in Section 4.2. The iterative 2DLDA method with the VM measure still performs poorly on this face database. The performance of the one-sided 2DLDA method with the volume measure is competitive with that of the one-sided 2DLDA method with the Frobenius measure. There is no clear winner between the volume measure and the Frobenius measure for the one-sided 2DLDA method. It is also noted that 2DLDA with the AMD measure is superior to the corresponding 2DLDA method with the Frobenius measure since the AMD measure contains the Frobenius measure. Likewise, LDA with the p-norm measure is superior to LDA with the ED measure since the p-norm measure contains the ED measure. The experiments further show that choosing distance measures for feature matrices is important for the 2DLDA method and the 2DLDA method is not always superior to the LDA method in real-world applications.

5. Conclusions

Although matrix-based discriminant methods have gained much attention recently, there still exist some relations between vector-based discriminant methods and matrix-based discriminant methods to be solved. In this paper, we carry out the theoretical and experimental analysis between them and explore the relationship between them. It is found that the discriminant power of 2DLDA is weaker than that of LDA when the same reduction dimension is considered. We also find that LDA outperforms 2DLDA in the large sample size problem. In the small sample size problem, the performance of LDA which considers the null space of the within-class scatter matrix is competitive with that of 2DLDA even if the number of training samples for each class is small in some cases. Experimental results also show that different distance measures in matrix form affect the performance of 2DLDA. Overall, 2DLDA does not always achieve better classification results than LDA when the number of training samples per class is small in real-world applications.

Acknowledgements

The authors would like to thank the reviewers for the constructive advice and comments. This work was partially supported by the Research Grants of Council of Hong Kong (No. CityU117106) and the National Natural Science Foundation of PR China (No.50775009).

References

Belhumeur, P.N., Hespanda, J., Kriegeman, D., 1997. Eigenfaces vs Fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Machine Intell. 19 (7), 711–720.

Blake, C., Keogh, E., Merz, C.J., 1998. UCI repository of machine learning databases. University of California, Irvine, CA <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

Cho, D., Chang, U., Kim, K., Kim, B., Lee, S., 2006. (2D)2DLDA for efficient face recognition. LNCS 4319, 314–321.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, second ed. Academic Press, Boston, MA.

Howland, P., Park, H., 2004. Generalized discriminant analysis using the generalized singular value decomposition. IEEE Trans. Pattern Anal. Machine Intell. 8, 995–1006.

Kongsontana, S., Rangsanseri, Y., 2005. Face recognition using 2DLDA algorithm. In: Proc. 8th Internat. Symp. Signal Process. Appl., pp. 675–678.

Kong, H., Wang, L., Teoh, E.K., 2005. A framework of 2D fisher discriminant: Application to face recognition with small number of training samples. In: Internat. Conf. CVPR.

Li, M., Yuan, B., 2005. 2D-LDA: A novel statistical linear discriminant analysis for image matrix. Pattern Recognition Lett. 26 (55), 527–532.

Noushath, S., Kumar, G., Shivakumara, P., 2006. (2D)LDA: An efficient approach for face recognition. Pattern Recognition 39, 1396–1400.

Meng, J., Zhang, W., 2007. Volume measure in 2DPCA-based face recognition. Pattern Recognition Lett. 28 (10), 1203–1208.

ORL, 1992. The ORL face database at the AT&T (Olivetti) research laboratory. Available from: <http://www.uk.research.att.com/facedatabase.html>.

Wang, Guisong, Wu, X., Jia, Z., 2003. Matrix Inequality, second ed. <www.sciencep.com>.

Wang, L., Wang, Xiao, Feng, Jufu., 2006. On image matrix based feature extraction algorithms. IEEE Trans. Systems Man Cybernet. Part B, 194–197.

Xiong, H., Swamy, M.N.S., Ahmad, M.O., 2005. Two-dimensional FLD for face recognition. Pattern Recognition 38 (7), 1121–1124.

Yale database, 1997. Available from: <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>.

Yang, J., Yang, J.Y., Frangi, A.F., Zhang, D., 2003a. Uncorrelated projection discriminant analysis and its application to face image feature extraction. Internat. J. Pattern Recognition Artificial Intell. 17 (8), 1325–1347.

Yang, J., Zhang, D., Frangi, A.F., Yang, J., 2004b. Two-dimensional PCA: A new approach to appearance based face representation and recognition. IEEE Trans. Pattern Anal. Machine Intell. 26, 131–137.

Yang, J., Zhang, D., Xu, Y., Yang, J., 2005c. Two-dimensional discriminant transform for face recognition. Pattern Recognition, 1125–1129.

Ye, J., Janardan, R., Li, Q., 2004a. Two-dimensional linear discriminant analysis. Adv. Neural Information Process. Systems.

Ye, Jieping, Janardan, R., Park, C.H., Park, H., 2004b. An optimization criterion for generalized discriminant analysis on undersampled problems. IEEE Trans. Pattern Anal. Machine Intell. 8, 982–994.

Ye, J., 2005a. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. Machine Learning Res., 483–502.
Ye, J., Li, Q., 2005b. A two-stage discriminant analysis via QR decomposition. IEEE Trans. Pattern Anal. Machine Intell., 929–941.
Ye, J., 2005c. Generalized low rank approximation of matrices. Machine Learning 61, 167–191.
Zuo, M., Zhang, D., Wang, K., 2006. An assembled matrix distance metric for 2DPCA-based image recognition. Pattern Recognition Lett. 27 (3), 210–216.