Computing and Informatics, Vol. 30, 2011, 965–986

TWO-DIMENSIONAL HETEROSCEDASTIC FEATURE EXTRACTION TECHNIQUE FOR FACE RECOGNITION

Mehran Safayani, Mohammad Taghi Manzuri Shalmani

Department of Computer Engineering
Sharif University of Technology, Tehran, Iran
e-mail: safayani@ce.sharif.edu, manzuri@sharif.edu

Communicated by Steve J. Maybank

Abstract. One limitation of vector-based LDA and its matrix-based extension is that they cannot deal with heteroscedastic data. In this paper, we present a novel two-dimensional feature extraction technique for face recognition which is capable of handling the heteroscedastic data in the dataset. The technique is a general form of two-dimensional linear discriminant analysis. It generalizes the interclass scatter matrix of two-dimensional LDA by applying the Chernoff distance as a measure of separation of every pair of clusters with the same index in different classes. By employing the new distance, our method can capture the discriminatory information presented in the difference of covariance matrices of different clusters in the datasets while preserving the computational simplicity of eigenvalue-based techniques. So our approach is a proper technique for high-dimensional applications such as face recognition. Experimental results on CMU-PIE, AR and AT & T face databases demonstrate the effectiveness of our method in term of classification accuracy.

Keywords: Linear discriminant analysis, heteroscedastic LDA, dimension reduction, feature extraction, face recognition, subspace learning

1 INTRODUCTION

Recently, feature extraction for face recognition in subspace domain has attracted growing attention. Principal component analysis (PCA) [1–3] and Linear discriminant analysis (LDA) [4–6] are two well-known methods which the former maintains the global Euclidean structure of the data in the original high dimensional space and

the latter preserves discriminative information between data of different classes. One limitation of classical LDA is that it implicitly assumes that the intraclass covariance matrices are identical. In other words, LDA ignores the discriminative information while the covariance matrices of different classes are not identical. Hence, it cannot deal with heteroscedastic data.

For solving this problem, different extensions of LDA have been proposed. Campbell derived a maximum likelihood approach for estimating the parameters of LDA with assuming that class covariances are identical and class means lie in a low dimensional subspace [7]. Following of Campbell, Kumar and Andreou proposed an iterative algorithm called Heteroscedastic Discriminant Analysis (HDA) by dropping the assumption of equal class covariances [8]. Loog and Duin introduced the Chernoff criterion for extending the LDA and employed the eigenvalues decomposition for computing the optimal projection matrix [9]. Rueda and Herrera proposed a linear dimension reduction which maximized the Chernoff distance in the transformed space [10]. Das and Nenadic suggested a criterion based on an approximation of an information-theoretic measure for handling the heteroscedastic data [11].

All the aforementioned LDA based methods and their heteroscedastic extensions convert an image matrix into a vector by concatenating its rows or columns. They do not consider the spatial information existed in the image and assume each pixel as an independent piece of information. Therefore, these methods involve the eigenvalue decomposition of huge matrices, which is very time-consuming. Also, the large number of parameters and the small number of training samples lead to the small sample size problem. Since the heteroscedastic solution considers individual covariance matrix for each class, its computational cost is much more than LDA. In addition, the small sample size problem is intensified in that the number of parameters needed to be estimated is increased. As a result, heteroscedastic extensions of LDA cannot directly be applied in high dimensional applications such as face recognition.

Another line of research in feature extraction is the matrix-based approaches. In these methods, the scatter matrices are constructed from the original image matrices. 2DPCA [12] and 2DLDA [13] are two well-known matrix-based algorithms constructed based on this idea. Some other researchers applied multilinear algebra and extended this concept to higher-order tensor data [14–17]. In general, these approaches not only reduce the computational cost by decreasing the number of projection parameters to be learned, but also preserve some implicit structural or locally-spatial information among elements of the original images. They also overcome the singularity problem of scatter matrices resulting from the high dimensionality of vectors.

To the best of our knowledge, in spite of particular interest in the matrix-based approaches but few papers investigate the heteroscedastic problem of them. Recently, some heteroscedastic extensions of 2DLDA have been proposed [18, 19]. In these works, an objective function that specifies an individual covariance matrix for each class is defined and then a numerical optimization routine is adapted for

finding the optimal transformation. However, in Section 3, we show that these approaches cannot solve this problem completely. On the other hand, Zheng et al. have investigated the bayes optimality conditions of Unilateral Two-Dimensional LDA (U2DLDA) and have expressed that heteroscedastic problem exists in U2DLDA [20]. However, they have not proposed any solution for it.

In this paper, we introduce a new discriminant feature extraction technique called Two-Dimensional Heteroscedastic Discriminant Analysis (2DHDA) which can handle the heteroscedastic data.

In each iteration of 2DHDA, one projection direction is fixed and the samples are projected to this projection matrix and then each column vector of the projected image matrices are considered as a new object and the cluster-based discriminant analysis is applied by clustering these samples according to their column indices. Interclass scatter matrix of our approach can capture the discriminatory information that is present in the difference of covariance matrices of different clusters in the dataset by using the Chernoff criterion. We also express that our proposed method is a general form of 2DLDA and reduces to it by assuming the identical covariance matrix for each cluster in the dataset.

Experimental results on three face databases denote that our proposed method is superior to the traditional vector-based and matrix-based approaches and their heteroscedastic extensions in term of classification accuracy.

The remaining part of the paper is organized as follows: Heteroscedastic problem is described in Section 2. Section 3 introduces our algorithm. We report the experimental results on the classification accuracy in Section 4. Finally, conclusions are brought in Section 5.

2 HETEROSCEDASTIC PROBLEM

Some of the important notations used in the rest of this paper are listed in Table 1.

LDA estimates the within class covariance matrix by averaging the individual class covariances as:

$$G_w = \frac{1}{N}\sum_{i=1}^{C} p_c G_{w,i}, \tag{1}$$

where p_i is the priori probability of i^{th} class, N and C are the total number of samples and classes, respectively. $G_{w,i}$, the covariance matrix of i^{th} class, is defined as

$$G_{w,i} = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_j - \bar{x_i})(x_j - \bar{x_i})^T, \tag{2}$$

where n_i denotes the total number of samples in the i^{th} class, x_j is the j^{th} sample vector from the i^{th} class and $\bar{x_i}$ is the average vector of the samples belonging to class i [21]. It is known that LDA is bayes optimal when the classes have normal distribution with identical covariance matrices. Heteroscedastic extension of LDA removes this assumption with assuming unequal covariance matrix for each class [8].

Notation	Description
N	number of images in the dataset
C	number of classes in the dataset
n_i	number of images in the i^{th} class
l_1	reduced dimensionality corresponding to the rows
l_2	reduced dimensionality corresponding to the columns
p_i	priori probability of i^{th} class
π_i	$p_i/(p_i+p_j)$, relative prior
t	iteration number
α	regularization parameter
x_i	i^{th} image in vectorized representation
X_i	i^{th} image in matrix representation
r	number of rows in X_i
c	number of columns in X_i
L	transformation matrix (left) by 2DLDA
R	transformation matrix (right) by 2DLDA
Y_i^R	i^{th} matrix image projected onto R transformation matrix
Y_i^L	i^{th} matrix image projected onto L transformation matrix
$tr(B)$	trace of B
$y_i^{R,j}$	j^{th} column of matrix Y_i^R
G_w	within-class scatter matrix
G_b	between-class scatter matrix
$G_{w,i}$	the covariance matrix of i^{th} class
G_w^R	within-class scatter matrix of right-projected image matrices
G_b^R	between-class scatter matrix of right-projected image matrices
$G_{w,c}^R$	sample covariance of right-projected image matrices of c^{th} class
$G_{w,c}^{R,j}$	sample covariance of j^{th} column of right-projected image of c^{th} class

Table 1. The notations of variables in the paper.

Recently, it has been shown that one important condition for Bayes optimality of 2DLDA is that the covariance matrices of the columns with the same index of image matrices within each class be identical [20]. Therefore, 2DLDA cannot handle the discriminant information that is present in the difference of covariances. In the other word this method cannot deal with heteroscedastic data.

The objective function of 2DLDA is

$$(R^*, L^*) = \underset{R,L}{\operatorname{argmax}} \frac{\sum_{i=1}^{C} p_i \|L^T \bar{X}_i R - L^T \bar{X} R\|^2}{\sum_{j=1}^{N} \|L^T X_j R - L^T \bar{X}_{c_j} R\|^2}, \tag{3}$$

where $X_j \in \Re^{(r\times c)}$ is the j^{th} sample image matrix of resolution $r \times c$ in the dataset. $\bar{X}_i$ is the average matrix of the samples belonging to class i, $\bar{X}$ is the average matrix over all the training samples, and the class label of X_j is c_j. There is not a closed-form solution for Equation (3), so an iterative algorithm for finding the local

optimal projections was proposed [13]. In each iteration, one projection direction like $R \in \Re^{(c \times l_2)}$ is assumed known, and the image samples are projected onto this projection matrix as follows:

$$Y_i^R = X_i R, \tag{4}$$

so the optimization problem in Equation (3) can be reformulated as a following discriminant analysis:

$$L^* = \underset{L}{\operatorname{argmax}} \frac{tr\left(L^T G_b^R L\right)}{tr\left(L^T G_w^R L\right)}, \tag{5}$$

where G_b^R and G_w^R are the right interclass and intraclass scatter matrix, respectively and are defined as:

$$G_b^R = \sum_{c=1}^{C} \sum_{j=1}^{l_2} p_c \left(\bar{y}_c^{R,j} - \bar{y}^{R,j}\right) \left(\bar{y}_c^{R,j} - \bar{y}^{R,j}\right)^T, \tag{6}$$

$$G_w^R = \sum_{i=1}^{N} \sum_{j=1}^{l_2} p_i \left(y_i^{R,j} - \bar{y}_{c_i}^{R,j}\right) \left(y_i^{R,j} - \bar{y}_{c_i}^{R,j}\right)^T, \tag{7}$$

where $y_i^{R,j}$represents the j^{th} column vector of the Y_i^R which is the right projected matrix from the sample matrix X_i. $\bar{y}_c^{R,j}$ is defined in the same way as $y_i^{R,j}$ with respect to the matrix $\bar{X}_c^R$, and l_2 denotes the number of column vectors of matrix y_i^R. It is known that, the optimization problem in Equation (5) is a special cluster-based discriminant analysis where the column vectors of the Y_i^R are considered as the new objects with the same class label as the original sample matrix and are clustered according to their column indices [22]. The optimal L is obtained by solving the following generalized eigenvalue problem: $G_b^R L = G_w^R L \Lambda_L$. Similarly, with the computed L, the G_w^L and $G_{w,c}^L$ are computed and the optimal R is obtained by solving another optimization problem $G_b^L R = G_w^L R \Lambda_R$.

For showing the heteroscedastic problem of 2DLDA, we reformulate Equation (7) as follows:

$$G_w^R = \sum_{c=1}^{C} p_c G_{w,c}^R, \tag{8}$$

where $G_{w,c}^R$, the right covariance matrix of c^{th} class, is defined as

$$G_{w,c}^R = \sum_{j=1}^{l_2} G_{w,c}^{R,j}, \tag{9}$$

where $G_{w,c}^{R,j}$, the right covariance matrix of j^{th} cluster in the c^{th} class, is defined as

$$G_{w,c}^{R,j} = \frac{1}{n_c} \sum_{i=1}^{n_c} \left(y_{c,i}^{R,j} - \bar{y}_c^{R,j}\right) \left(y_{c,i}^{R,j} - \bar{y}_c^{R,j}\right)^T. \tag{10}$$

As can be seen from the Equations (8) and (9), there are two plug-in estimates for computing the G_w^R. First, $G_{w,c}^R$, the c^{th} class covariance matrix, is estimated by the average of within-cluster covariance matrices, i.e., $G_{w,c}^{R,j}$'s, then the intraclass covariance is estimated using the individual class covariances.

If the clusters within each class are heteroscedastic, i.e. $G_{w,c}^{R,i} \neq G_{w,c}^{R,j}$ for $i \neq j$, the first estimate becomes improper. Since, in face datasets, the distribution of clusters, i.e. columns of image with different indices, is substantially different, so heteroscedastic problem should be addressed. The other estimation is similar to that performing in classic LDA in term of Equation (1), i.e., estimating the intraclass covariance matrix from the individual class sample-covariances, which may fail due to the unequal class covariance matrices. This problem also can be shown in computing of left covariance matrices G_w^L. Therefore, we can conclude that 2DLDA suffers from heteroscedastic problem and this problem is more serious than vector-based LDA.

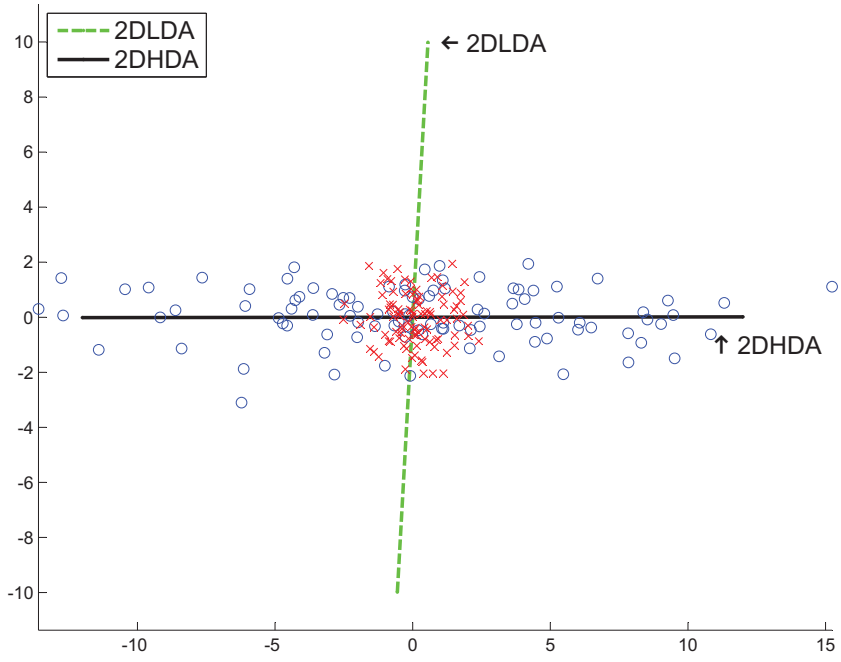

Fig. 1. A toy example of two-class pattern classification problem

Figure 1 illustrates the heteroscedastic problem of 2DLDA with synthetic data. The data in two dimensions, belonging to two classes shown with cross and circle, is to be projected onto a single dimension. The cross class is generated from a Gaussian

distribution with $\mu_1 = (0, 0)$ and $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ while circle class comes from a Gaussian with $\mu_2 = (0, 0)$ and $\Sigma_2 = \begin{pmatrix} 30 & 0 \\ 0 & 1 \end{pmatrix}$. From this toy example, it can be observed that 2DLDA cannot get the proper direction when two classes have the same mean and different covariance matrices. Therefore, 2DLDA cannot handle heteroscedasticity in the data, and it chooses the direction which leads to strong class overlap. This figure also shows that our proposed method, 2DHDA, takes account of the discriminant information in the difference of covariance matrices and leads to a projection which gives much better class separation.

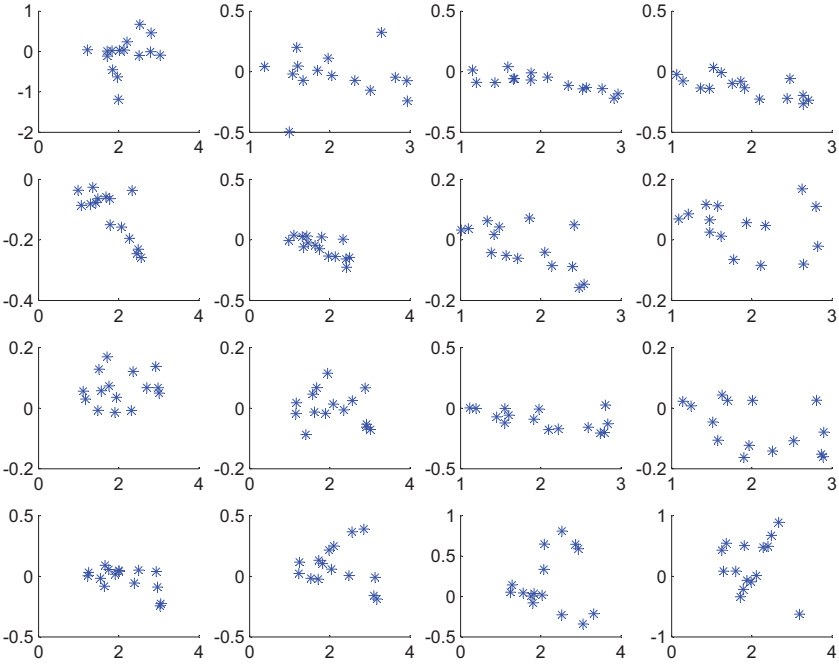

Fig. 2. Data distributions of different columns of the images belonging to a specific person in CMU-PIE database. Each subplot corresponds to one column index. The column index of the left top plot is 2 and index increases in order of left top to right down with step 2.

To get a better understanding of the heteroscedasticity in the face database, we run a test with real dataset. Data of the test comes from the PIE face database described in Section 4.1. We consider 65 persons each having 13 images with the resolution of 32×32 pixels. The columns of the images are projected onto the two leading eigenvectors generated using 2DPCA [12] algorithm, therefore each image

has 32 two-dimensional column vectors. Figure 2 presents that the column vectors of the images belonging to a specific person have different distributions. Each subplot in this figure depicts the column vectors correspond to one column index of the images ranging from 2 to 32 with step two.

It can be observed that distinction between distributions of the data in different subplots is very large. In other words, the distribution of various columns of face images in a class is significantly different. In this case, the heteroscedastic problem has to be addressed, and "plug-in" estimate in term of Equation (9) is improper.

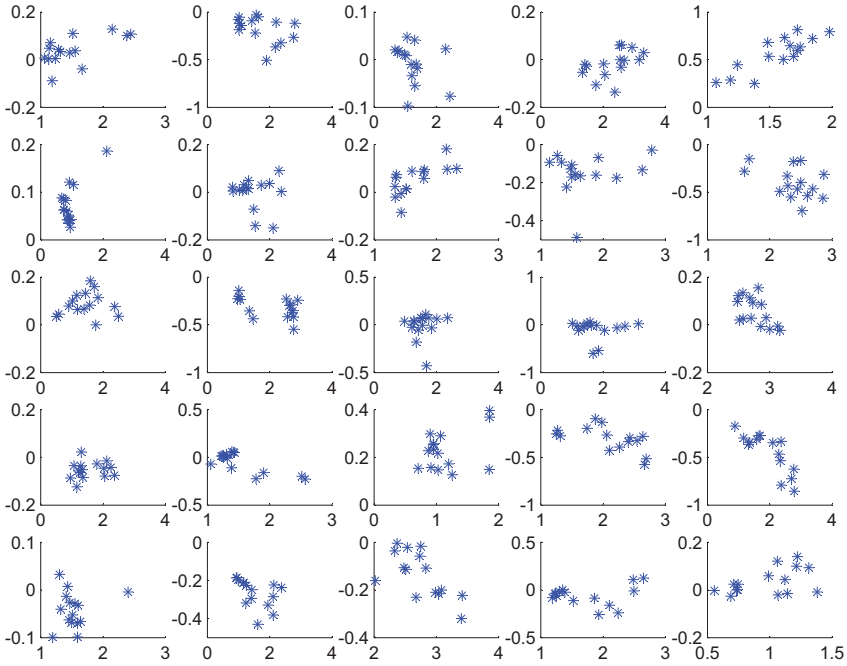

Fig. 3. Each plot illustrates data samples drawn from the one of the classes of PIE face database

For showing the heteroscedasticity in the different classes in the dataset, we select 25 out of 65 persons in the database. Then, for each class we only consider the column vectors with specific column index. Each subplot in Figure 3 illustrates the data samples corresponding to one of the classes in the dataset. This figure demonstrates that the distribution of the data with same column index in different classes are completely different and implicitly declares that estimating the within-class covariance matrix by averaging the within-class covariance matrix of individual

classes is inappropriate because the large difference of covariance matrix of different classes.

In general, these examples express that heteroscedastic problem exists in the face dataset, and also the estimates of within-class covariance matrix in two-dimensional LDA are not proper and degrades the performance of this method.

3 TWO-DIMENSIONAL HETEROSCEDASTIC DISCRIMINANT ANALYSIS (2DHDA)

LDA discards the discriminative information in the difference of covariance matrices, so it does not take into account the discriminative information contained in the covariances of different classes. For solving this problem, in [23], a criterion was defined containing the average of interclass divergences. This method considered the variations in the covariance matrices of different classes by maximizing this criterion. In [24], a dimensionality reduction for multimodal Gaussian classes with different covariances called Multimodal Oriented Discriminant Analysis (MODA) was derived. In [7], a relationship between LDA and estimated parameters of a Gaussian model using maximum likelihood is established. Heteroscedastic Discriminant Analysis (HDA) was proposed in [8] by dropping the identical class covariances assumption. In [9], directed distance metric was introduced and an extension of LDA based on the Chernoff criteria was developed. In [25], a method called General Averaged Divergence Analysis (GADA) based on the generalized Kullback-Leibler divergence between different classes was proposed. In [10], a linear dimension reduction technique was presented which aims to maximize the Chernoff distance in the transformed space, and thus increases the class separability in such a space. In [11], a method was proposed that maximizes a criterion that belongs to the class of probability dependence measures, and is naturally defined for multiple classes. The criterion is based on an approximation of an information-theoretic measure and is capable of handling heteroscedastic data.

In the previous section, it has been shown that 2DLDA has heteroscedastic problem and this problem is different from that of LDA. The existing reported matrix-based heteroscedastic LDA's add 2D constraint on the traditional vector-based Heteroscedastic LDA (HLDA) by constructing the covariance matrices using the original image matrices [18, 19]. Similar to HLDA, they remove the restriction that all the within-class covariance matrices are the same. Therefore, they only consider the estimate in terms of Equation (8) and cannot deal with heteroscedasticity in the columns of the image matrices which can degrade the recognition performance. Also, they adopt only one subspace, and the disadvantage arising in this way is that more coefficients are needed to represent an image in these methods than in HLDA.

In our approach, the column vectors of the sample matrices are clustered according to their column indices, and the individual covariance matrix for each cluster is specified. Then, the distance between every pair of clusters with the same index of column in the different classes is maximized by using the Chernoff distance measure.

Chernoff distance can be used as a measure of separability of two distributions. the distance for normal distributions is defined as follows [26]:

$$D_c = \frac{s(1-s)}{2}(\mu_2-\mu_1)^T[s\Sigma_1+(1-s)\Sigma_2]^{-1}(\mu_2-\mu_1)+\frac{1}{2}\ln\frac{|s\Sigma_1+(1-s)\Sigma_2|}{|\Sigma_1|^s|\Sigma_2|^{(1-s)}} \quad (11)$$

where μ_i and Σ_i are the mean vector and covariance matrix of the i^{th} class, respectively and s is a constant in the range $[0,1]$. As can be observed from Equation (11), this distance can capture differences in within-class covariance matrices and the discriminatory information therein, so it can make better use of the information in heteroscedastic data.

We start with 2DLDA objective function in two-class case and then generalized it so that it can capture the difference in the covariances.

3.1 Two-Class Case and $R \in \Re^{(c\times 1)}$

We assume that $R \in \Re^{(c\times 1)}$, the right projection matrix, contains only one eigenvector corresponding to the leading eigenvalue, and we also assume that $G_w^R = I$, where I is the identity matrix. Therefore, in this case, regarding to Equations (4) and (5), we have

$$L^* = \underset{L}{\operatorname{argmax}} \frac{tr\left(L^T G_E^R L\right)}{tr\left(L^T L\right)}, \quad (12)$$

where $G_E^R = \left(\bar{y}_1^R - \bar{y}_2^R\right)\left(\bar{y}_1^R - \bar{y}_2^R\right)^T$. This criterion is maximized by the eigenvalue decomposition of G_E^R which only has one none zero eigenvalue equal to the trace of G_E^R. This eigenvalue expresses the square Euclidean distance between the mean of the two classes. For handling heteroscedasticity of the data and keeping more discriminatory information, we replace G_E^R by G_C^R whose trace is equal to the Chernoff distance [26] between two class distributions.

$$G_C^R = \left(G_w^R\right)^{-1/2} G_E^R \left(G_w^R\right)^{-1/2} + \frac{1}{p_1 p_2}\left(\log G_w^R - p_1 \log G_{w,1}^R - p_2 \log G_{w,2}^R\right), \quad (13)$$

where $\log(A)$ is defined as $R(\log(V))R^{-1}$, and RVR^{-1} is the eigenvalue decomposition of A where A is a symmetric positive definite matrix.

If $G_w^R \neq I$, the data are sphered using $\left(G_w^R\right)^{-1/2}$ transformation. In the transformed domain, $\bar{Y}_j^R$, $G_{w,i}^R$ and G_w^R become $\left(G_w^R\right)^{-1/2}\bar{Y}_j^R$, $\left(G_w^R\right)^{-1/2} G_{w,i}^R \left(G_w^R\right)^{-1/2}$, and I, respectively. Then, G_C^R is computed, and the inverse transform $\left(G_w^R\right)^{1/2}$ is applied to get the matrix $\ddot{G}_C^R = \left(G_w^R\right)^{1/2} G_C^R \left(G_w^R\right)^{1/2}$, which can be rewritten as:

$$\ddot{G}_C^R = G_E^R + \frac{1}{p_1 p_2}\left(G_w^R\right)^{1/2}\left(\log G_w^R - p_1 \log G_{w,1}^R - p_2 \log G_{w,2}^R\right)\left(G_w^R\right)^{1/2}. \quad (14)$$

3.2 Two-Class Case and $R \in \Re^{c \times l_2}$

In this case, G_B^R, the interclass scatter matrix of 2DLDA, can be rewritten as follows:

$$G_B^R = \sum_{s=1}^{l_2} p_1 p_2 \left(\bar{y}_1^{R,s} - \bar{y}_2^{R,s}\right)\left(\bar{y}_1^{R,s} - \bar{y}_2^{R,s}\right)^T = \sum_{s=1}^{l_2} p_1 p_2 G_E^{R,s}, \tag{15}$$

where $G_E^{R,s} = \left(\bar{y}_1^{R,s} - \bar{y}_2^{R,s}\right)\left(\bar{y}_1^{R,s} - \bar{y}_2^{R,s}\right)^T$ is the scatter matrix which captures the difference between the mean vectors of the s^{th} cluster of two classes. We generalize interclass scatter matrix by replacing $G_E^{R,s}$ with Chernoff scatter matrix $\dddot{G}_C^{R,s}$.

$$\dddot{G}_C^{R,s} = G_E^{R,s} + \frac{1}{p_1 p_2}\left(G_w^R\right)^{1/2}\left(\log G_w^{R,s} - p_1 \log G_{w,1}^{R,s} - p_2 \log G_{w,2}^{R,s}\right)\left(G_w^R\right)^{1/2}. \tag{16}$$

3.3 Multi-Class Case and $R \in \Re^{c \times l_2}$

In Sections 3.1 and 3.2, we express the formulation for two-class case. In this section, we generalize it to multi-class case. The interclass scatter matrix can be decomposed to:

$$G_B^R = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\sum_{s=1}^{l_2} p_i p_j \left(\bar{y}_i^{R,s} - \bar{y}_j^{R,s}\right)\left(\bar{y}_i^{R,s} - \bar{y}_j^{R,s}\right)^T = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\sum_{s=1}^{l_2} p_i p_j G_{E,i,j}^{R,s}. \tag{17}$$

This formula can be generalized by replacing $G_{E,i,j}^{R,s}$ by $G_{C,i,j}^{k,s}$ which is the scatter matrix which captures the Chernoff distance between the s^{th} cluster of class i and j.

$$\dddot{G}_{C,i,j}^{R,s} = G_{E,i,j}^{R,s} + \frac{1}{\pi_i \pi_j}\left(G_{w,i,j}^{R,s}\right)^{1/2}\left(\log G_{w,i,j}^{R,s} - \pi_i \log G_{w,i}^{R,s} - \pi_j \log G_{w,j}^{R,s}\right)\left(G_{w,i,j}^{R,s}\right)^{1/2}, \tag{18}$$

where $\pi_i = p_i/(p_i + p_j)$, $\pi_j = p_j/(p_i + p_j)$ are relative priors, and $G_{w,i,j}^{R,s} = \pi_i G_{W,i}^{R,s} + \pi_j G_{W,j}^{R,s}$. Therefore, $\dddot{G}_C^R$ obtains as follows:

$$\dddot{G}_C^R = \sum_{i=1}^{C}\sum_{j=i+1}^{C-1}\sum_{s=1}^{l_2} p_i p_j \dddot{G}_{C,i,j}^{R,s}, \tag{19}$$

then, the new optimization formula becomes:

$$L^* = \underset{L}{\operatorname{argmax}} \frac{tr\left(L^T \dddot{G}_C^R L\right)}{tr\left(L^T G_w^R L\right)}. \tag{20}$$

This objective function can be solved by the generalized eigenvalue decomposition method as follows:

$$\dddot{G}_C^R L = G_W^R L \Lambda_L, \tag{21}$$

where Λ_L is diagonal matrix whose entities are eigenvalues of $(G_w^R)^{-1}\dddot{G}_C^R$, sorted in the descending order, and L is a matrix whose columns are the corresponding eigenvectors. Similarly, with computed L, the optimal projection matrix R is computed using the following general eigenvalue decomposition

$$\dddot{G}_C^L R = G_W^L R \Lambda_R, \tag{22}$$

where G_w^L and $\dddot{G}_C^L$ are defined as follows:

$$G_w^L = \sum_{c=1}^{C} p_c G_{w,c}^L, \tag{23}$$

$$\dddot{G}_C^L = \sum_{i=1}^{C} \sum_{j=i+1}^{C-1} \sum_{s=1}^{l_1} p_i p_j \dddot{G}_{C,i,j}^{L,s}. \tag{24}$$

Summarize procedure of 2DHDA is given in Figure 4.

Input: the sample set, $X_i \in \Re^{r\times c}$, $i = 1, \ldots, N$,
class label $c_i \in \{1, 2, \ldots, C\}$, and
the final lower dimensions $l_1 \times l_2$
Output: Find $L \in \Re^{r\times l_1}$, $R \in \Re^{c\times l_2}$
Initialize: $R^0 = I_{c\times c}$
for $t = 1, \ldots, T_{max}$
  $Y_j^R = X_j R^{t-1}$, $j = 1, \ldots, N$
  Compute G_w^R from Equation (8)
  $Y_j^R = (G_W^R)^{-1/2} Y_j^R$
  Compute $\dddot{G}_C^R$ from Equation (19)
  $L^t \dddot{G}_C^R = G_w^R L^t \Lambda_L$, $L^t \in \Re^{r\times l_1}$
  $Y_j^L = (L^t)^T X_j$, $j = 1, \ldots, N$
  Compute G_w^L from Equation (23)
  $Y_j^L = (G_W^L)^{-1/2} Y_j^L$
  Compute $\dddot{G}_C^L$ from Equation (24)
  $R^t \dddot{G}_C^L = G_w^L R^t \Lambda_R$, $R^t \in \Re^{c\times l_2}$
end.

Fig. 4. 2DHDA procedure

3.4 Connection to 2DLDA

Theorem 1. If all the covariance matrices are the same, (i.e. $G_i^{R,s} = \Sigma_R \forall i, s$) then 2DHDA objective function is equal to that of 2DLDA.

Proof. Without loss of generality we prove the equivalence for right projection matrix. If $G_i^{R,s} = \Sigma_R \; \forall i, s$ then $\ddot{G}_{w,i,j}^{R,s} = \Sigma_R$ and in term of Equation (18) we have

$$\dddot{G}_{C,i,j}^{R,s} = G_{E,i,j}^{R,s} + \frac{1}{\pi_i \pi_j} (\Sigma_R)^{1/2} \left((1 - (\pi_i + \pi_j)) \log \Sigma_R \right) (\Sigma_R)^{1/2}, \tag{25}$$

where $(\pi_i + \pi_j) = 1$, therefore $\dddot{G}_{C,i,j}^{R,s} = G_{E,i,j}^{R,s}$ and consequently according to Equations (17) and (19), $\dddot{G}_C^R$ is equal to G_B^R. □

3.5 Regularization

Our approach can be enhanced by smoothing the covariance matrices by adding different kind of regularization techniques [21]. Here, for avoiding the singular matrices, we add small regularization term of α to the within class covariance when it is necessary.

$$\hat{G}_{w,i,j}^{R,s} = \dddot{G}_{w,i,j}^{R,s} + \alpha I_r, \tag{26}$$

where I_r is the identity matrix of size $r \times r$.

4 EXPERIMENTS

In this section, we investigate performance of our proposed subspace learning approach for face recognition. The recognition has three steps: first, face subspace is computed using the training set. Then, test images are projected onto the low dimensional subspaces. Finally the test images are identified using nearest neighbor classifier. In all experiments each image is manually cropped and resized to 32×32 pixels, with 256 gray levels per pixel. The pixel values of each image is normalized to $[0, 1]$, and the resulting image is preprocessed using a histogram-equalization.

We randomly select two images per subject for training and the rest for testing. The experiments are repeated 20 times with different groups of training images, and the mean as well as standard deviation of the results are reported.

4.1 Datasets

The CMU PIE face database contains 68 subjects with 41,368 face images as a whole [27]. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. The subset "CMU-PIE" is established by selecting images under natural illumination for all persons from the frontal view,1/4 left/right profile and below/above in frontal view (C05, C07, C09, C27, C29). For each view, there are three different expressions, namely natural expression, smiling and blinking. Hence there are 15 face images for each subject. Figure 5 shows 15 image samples of one person.

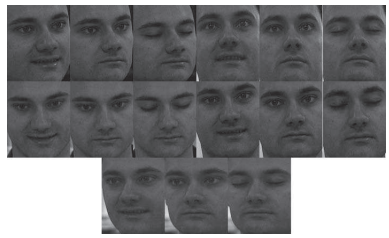

Fig. 5. 15 sample images of one person from CMU-PIE database

The AR face database contains over 3,200 frontal face images of 126 different individuals (70 men and 56 women) [28]. Most individuals have 26 different images taken in two different sessions separated by two weeks intervals and each session consists of 13 faces with different facial expressions, illumination conditions and occlusions. In our experiments, we use a subset of the AR face database which contains 650 face images corresponding to 50 persons (25 men and 25 women), where each person has 13 different images from the first section. Some examples from this database are illustrated in Figure 6.

Fig. 6. 13 different images of one person from AR database

AT & T face database is the third database used in our experiments [29]. This database contains images of 40 individuals, each providing 10 different images. The pose, expression, and facial details variations are also included. The images are taken with a tolerance for some tilting and rotation of the face of up to 20 degrees. Moreover, there are also some variations in the scale of up to about 10 percent. Ten sample images of one person from the AT & T database are shown in Figure 7.

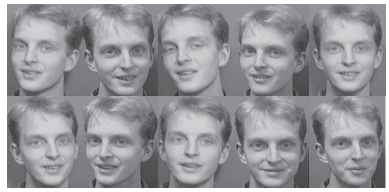

Fig. 7. Ten sample images of one subject in the AT & T face database

4.2 Compared Algorithms

We compare our method with five different subspace algorithms which belong to three different families, i.e., PCA family, LDA family and Heteroscedastic LDA family. In PCA family, we have vector-based PCA (Eigenface) [3] and its matrix extension, i.e., general low rank approximation of matrix (GLRAM) [30]. In LDA family, vector-based version of LDA, Fisherface [4] and its matrix-based version 2DLDA [13] have been investigated. In Heteroscedastic LDA family, Heteroscedastic Discriminant Analysis with Two-dimensional Constraints (HDA/2D) [19] has been compared. For HDA/2D, the objective function was optimized using the qusi-Newton numerical optimization of MATLAB toolbox. Projected features in the vector-based methods are expressed by a d-dimension vector and in the matrix-based approaches are shown by a $d_1 \times d_2$ dimensional matrix. For Fisherface maximum number of discriminant features is $C-1$, where C is the total number of subjects while that for eigenface is $N-1$ where N is the total number of subjects in the dataset. The number of features indicated for 2DLDA, HDA/2D, GLRAM and 2DHDA is the product of d_1 and d_2 dimension where for HDA/2D d_2 is 32.

4.3 Face Recognition Results

The top recognition accuracy of different algorithms on CMU-PIE, AR and AT & T databases are reported in Table 2. Specifically, each entry in the table shows the mean and standard deviation of recognition rate over 20 random splits. Also, the corresponding dimensions are listed in the parentheses on the right. For each dataset, the average recognition rate of an algorithm with respect to different numbers of discriminant features is computed. Based on these results we then express the best average recognition rates of the algorithm in the table.

Method	**CMU-PIE**	**AR**	**AT & T**
Eigenface	39.25 ± 5.49 (105)	44.36 ± 12.86 (95)	66.98 ± 3.94 (55)
GLRAM	54.79 ± 7.04 (11×11)	69.62 ± 7.81 (12×12)	79.36 ± 2.76 (10×10)
Fisherface	47.10 ± 7.17 (67)	63.62 ± 5.81 (49)	70.30 ± 3.95 (35)
2DLDA	46.99 ± 9.79 (6×6)	73.20 ± 4.80 (12×12)	78.50 ± 2.58 (8×8)
HDA/2D	52.66 ± 10.13 (10×32)	64.81 ± 6.33 (10×32)	77.72 ± 4.11 (10×32)
2DHDA	55.74 ± 7.84 **(5×5)**	76.25 ± 6.00 **(12×12)**	82.28 ± 2.92 **(5×5)**

Table 2. Face recognition accuracy of different subspace algorithms on CMU-PIE, AR and AT & T datasets [mean $\pm$ std % (dimension)]

Figure 8 illustrates the average recognition rate of each methods over different databases. It could be observed that 2DHDA outperforms all other methods on these datasets.

For further exploration, Figure 9 plots the face recognition rate with respect to the different number of features over CMU-PIE face database. As we can clearly

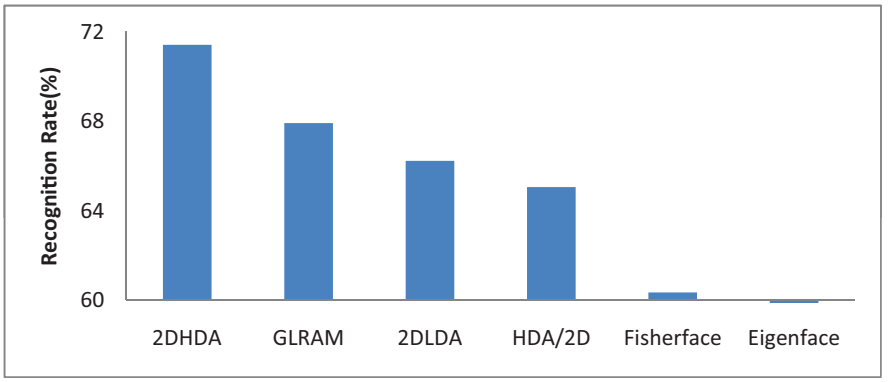

Fig. 8. Average recognition accuracy of different methods

observe, 2DLDA do not perform well on this database, possibly due to the heteroscedastic problem. 2DHDA achieves a recognition rate of 55.74 % with 25 features while followed by 2DLDA with a recognition rate of 46.99 % with 36 features. It is also observed that HDA/2D did not perform as well as 2DHDA method. GLRAM also has good performance. However, 2DHDA is the best method.

From Figure 9, it could be observed that the 2DLDA and 2DHDA 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.

This may occur as a result of using simple sum of Euclidean as the distance metric for the score of the match between two feature matrices obtained by two-dimensional methods. In our experiments, each column of the corresponding feature matrices is adopted with uniform weight, while more emphasis should be placed on the feature vectors corresponding to the larger eigenvalues.

In Figure 10, we present the recognition rate of different methods versus the number of discriminatory features over AR face database. Again, 2DHDA outperforms all other methods, which achieves a recognition rate of 76.25 % with 144 features. Despite the results of the previous experiment, in this experiment 2DLDA method performs better than GLRAM and other vector-based algorithms on AR dataset.

The recognition rate versus feature numbers of different methods is plotted once again over AT & T database in Figure 11. The recognition accuracy of our proposed method is 82.28 % with 25 features, which is better than all other methods. The 2DLDA method significantly outperforms other vector-based method while its performance is inferior to the GLRAM method.

Figure 12 shows the convergence characteristics of our algorithm. The horizontal axis indicates the number of iterations, and the vertical axis is the similarity of two successively estimated projection matrices, i.e. $tr(abs(U^t)^T * abs(U^{t-1}))$ where U could be left (L) or right (R) projection matrix and t denotes the iteration number. It is clearly observed that the 2DHDA method has no convergence problem.

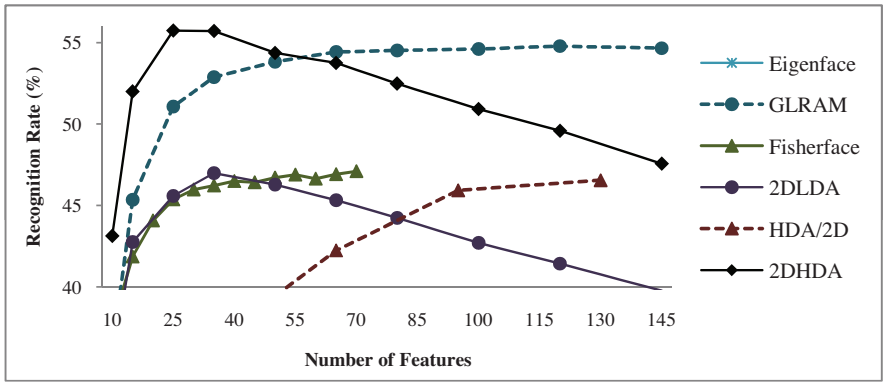

Fig. 9. Recognition accuracy versus different number of discriminant features on CMU-PIE dataset

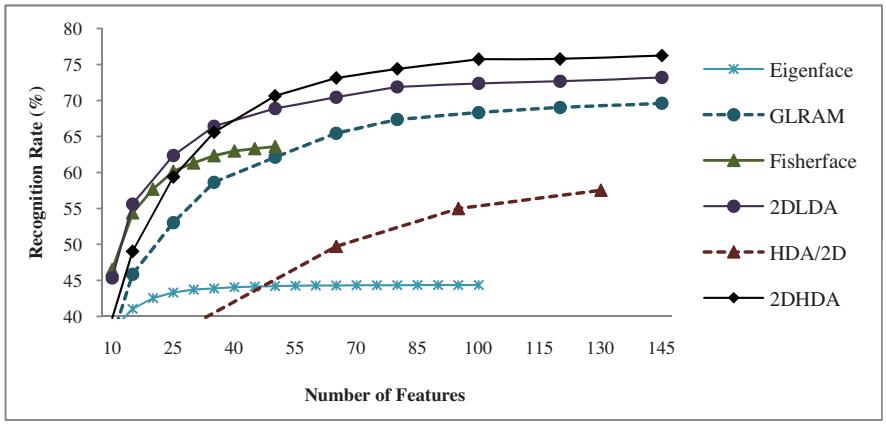

Fig. 10. Average recognition accuracy versus different number of discriminant features on AR dataset

Moreover, the similarity of the projection matrices of second and first iteration is about 90 %. Figure 13 illustrates the average recognition rate versus different number of iterations; it is clear that recognition rate is stable with respect to the number of iterations. The experiments also show that after one iteration the recognition accuracies are satisfying. Therefore, in our experiments, similar to [13], we employ our algorithm with only one iteration, which is less costly, ensures good recognition results, and frees us from determining the number of iterations.

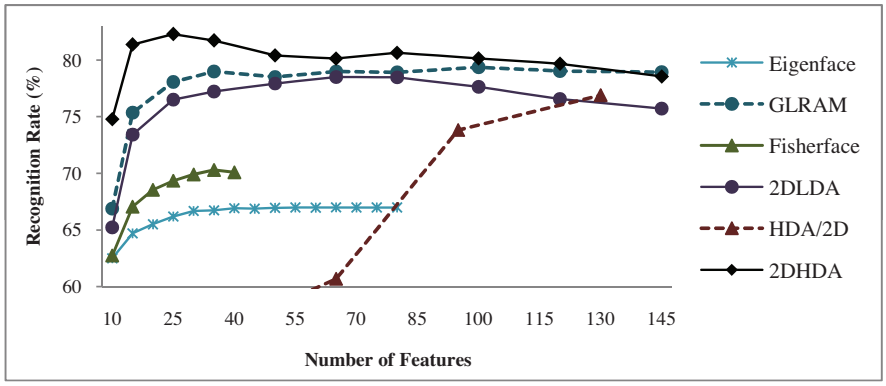

Fig. 11. Average recognition accuracy versus different number of discriminant features on AT & T dataset

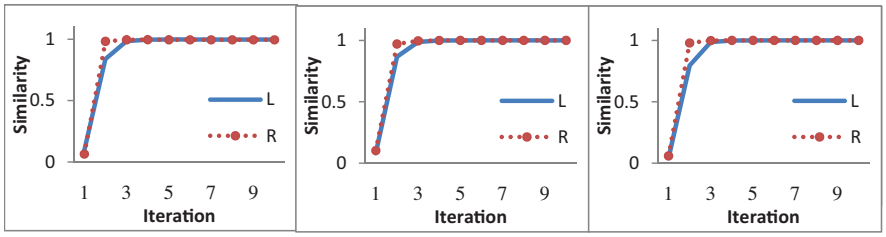

Fig. 12. Similarity of two successive estimated projection matrices of our proposed method versus iteration numbers on different datasets. Left: CMU-PIE, middle: AR, right: AT & T

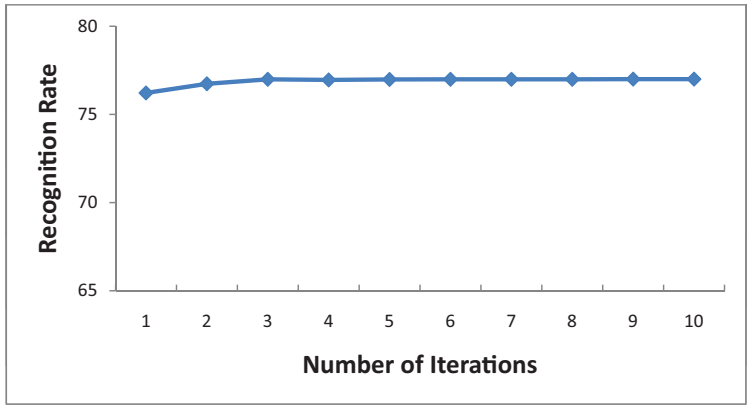

Fig. 13. Average recognition rate versus different number of iterations on AR face dataset

4.4 Discussion

The recognition rate of eigenface are quit low specially on CMU-PIE and AR, which indicates the difficulty of the databases. GLRAM which is a matrix-based method improves the recognition performance of vector-based eigenface. 2DLDA does not always yield higher accuracy than the vector-based LDA method. It outperforms Fisherface on AT & T and AR databases, while its performance descends on CMU-PIE dataset. HDA/2D has higher recognition rate than 2DLDA on CMU-PIE database, while its performance degrades on AR and AT & T database. It shows that defining the individual class covariances without regarding to the distribution of the clusters cannot necessarily improve the recognition performance of matrix-based approaches. 2DHDA always significantly outperform 2DLDA in different databases. The reason lies in relaxing the constraint of equality of the covariance matrices of different clusters.

Computational Cost: For ease of understanding, let us assume that the sample matrices has uniform numbers of rows and columns, i.e., $r = c = m$, and we set l_1 and l_2 to a common value d. Therefore, the complexity of 2DHDA is $O((N+C^2)dm^3)$ for each iteration.

5 CONCLUSION

In this paper a new approach called Two-Dimensional Heteroscedastic Discriminant Analysis (2DHDA) for solving the heteroscedastic problem of 2DLDA method was proposed. Our approach keeps the computational simplicity of eigenvalue-based techniques and benefits from the spatial redundancies in the image matrix, so it can be applied in high-dimensional application such as face recognition, where applying of vector-based heteroscedastic solutions is not feasible due to high computational cost.

We expressed that 2DLDA has two plug-in estimates and if the data of the columns with different indexes were heteroscedastic, then those estimations would be improper. It was shown that Euclidean distance between class means that 2DLDA cannot take into account difference in the intraclass covariance matrices and the discriminatory information within them.

For solving this problem the Chernoff distance was applied in computing of interclass covariance matrices. 2DHDA is also a general framework which can reduce to 2DLDA when all the intracluster covariance matrices are the same. Experiments on a number of datasets including CMU-PIE, AR and AT & T face databases demonstrate that 2DHDA consistently performs better across all the three datasets than 2DLDA and the other ordinary subspace learning algorithm.

Acknowledgement

This paper is partially supported by Iran Telecommunication Research Center (ITRC).

REFERENCES

[1] Joliffe, I.: Principal Component Analysis. Springer-Verlag 1986.

[2] Martinez, A.—Kak, A.: Pca versus lda. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 23, 2001, pp. 228–233.

[3] Turk, M.—Pentland, A.—Bsat, M.: Face Recognition Using Eigenfaces. Computer Vision and Pattern Recognition 1991, pp. 586–591.

[4] Belhumeur, P. N.—Hespanha, J.—Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19, 1997, pp. 711–720.

[5] Ye, J.—Janardan, C.—Park, C.—Park, H.: An Optimization Criterion for Generalized Discriminant Analysis on Undersampled Problems. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 26, 2004, pp. 982–994.

[6] Yu, H.—Yang, J.: A Direct Lda Algorithm for High-Dimensional Data With Application to Face Recognition. Pattern Recognition, Vol. 34, 2001, pp. 2067–2070.

[7] Campbell, N. A.: Canonical variate analysis a general formulation. Australian Journal of Statistics, Vol. 26, 1984, pp. 86–96.

[8] Kumar, N.—Andreou, A.: Heteroscedastic Discriminant Analysis and Reduced Rank HMMS for Improved Speech Recognition. Speech Communication, Vol. 26, 1998, pp. 283–297.

[9] Loog, M.—Duin, R.: Linear Dimensionality Reduction Via a Heteroscedastic Extension of lda: The Chernoff Criterion. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 26, 2004, pp. 732–739.

[10] Rueda, L.—Herrera, M.: Linear Dimensionality Reduction by Maximizing the Chernoff Distance in the Transformed Space. Pattern Recognition, Vol. 41, 2008, pp. 3138–3152.

[11] Das, K.—Nenadic, Z.: Approximate Information Discriminant Analysis: A Computationally Simple Heteroscedastic Feature Extraction Technique. Pattern Recognition, Vol. 41, 2008, pp. 1548–1557.

[12] Yang, J.—Zhang, D.—Frangi, A.—Yang, J.: Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 26, 2004, pp. 131–137.

[13] Ye, J.—Janardan, R.—Li, Q.: Two-Dimensional Linear Discriminant Analysis. In Proceedings of the Advances in Neural Information Processing Systems (NIPS) 2005, pp. 1569–1576.

[14] Cai, D.—He, X.—Han, J.: Subspace Learning Based on Tensor Analysis. Computer Science Department, University of Illinois at Urbana-Champaign (UIUC), Tech. Rep. UIUCDCS-R-2005–2572, 2005.

[15] He, X.—Cai, D.—Niyogi, P.: Tensor Subspace Analysis. In Proceedings of the Advances in Neural Information Processing Systems (NIPS) 2005, pp. 499–506.

[16] Yan, S.—Xu, D.—Zhang, B.—Zhang, H.: Graph Embedding: A General Framework for Dimensionality Reduction. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2005, pp. 830–837.

[17] Yan, S.—Xu, D.—Yang, Q.—Zhang, L.—Tang, X.—Zhang, H. J.: Multilinear Discriminant Analysis for Face Recognition. IEEE Trans. Image Processing, Vol. 16, 2007, pp. 212–220.

[18] Ueki, K.—Hayashida, T.—Kobayashi, T.: Two-Dimensional Heteroscedastic Linear Discriminant Analysis for Age-Group Classification. In Proceedings of the International Conference on Pattern Recognition (ICPR) 2006, pp. 585–588.

[19] Chen, S.—Yu, Y.—Luo, B.—Wang, R.: Heteroscedastic Discriminant Analysis With Two-Dimensional Constraints. In ICASSP 2008, pp. 4701–4704.

[20] Zheng, W.—Lai, J.—Li, S.: 1d-lda vs. 2d-lda: When Is Vector-Based Linear Discriminant Analysis Better Than Matrix-Based? Pattern Recognition, Vol. 41, 2008, pp. 2156–2172.

[21] Friedman, J. H.: Regularized Discriminant Analysis. Journal of the American Statistical Association, Vol. 84, 1989, pp. 165–175.

[22] Yan, S.—Xu, D.—Yang, Q.—Zhang, L.—Tang, X.—Zhang, H. J.: Discriminant Analysis With Tensor Representation. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2005, pp. 526–532.

[23] Decell, H. P.—Mayekar, S.: Feature Combinations and the Divergence Criterion. Computers and Math. With Applications, Vol. 3, 1977, pp. 71–76.

[24] De La Torre, F.—Kanade, T.: Multimodal Oriented Discriminant Analysis. In Proceeding of the International Conference on Machine Learning, Germany 2005.

[25] Tao, D.—Xuelong, L.—Xindong, W.—Maybank, S.: General Averaged Divergence Analysis. In Proceeding of Seventh IEEE International Conference on Data Mining (ICDM) 2007, pp. 302–311.

[26] Fukunaga, K.: Introduction to Statistical Pattern Recognition. 2nd ed., Computer Science and Scientific Computing Series, Academic Press 1990.

[27] Sim, T.—Baker, S.—Bsat, M.: The CMU Pose, Illumination, and Expression Database. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 25, 2003, pp. 1615–1618.

[28] Martinez, A.—Benavente, R.: The AR Face Database. Tech. Rep. CVC 24, 1998.

[29] AT & T Labratory Cambridge [Online]. Available on `http://www.cl.cam.ac.uk`.

[30] Ye, J.: General Low Rank Approximations of Matrices. Machine Learning, Vol. 61, 2005, pp. 167–191.

Mehran Safayani received the B. Sc. degree from Esfahan University, Esfahan, Iran, in 2002. After working at the same university for two years, he came to Sharif University of Technology, Tehran, Iran, and received his M. Sc. degree in Computer Engineering in 2006. He is currently a Ph. D. candidate at the Computer Science Department of this university. His research interests include pattern recognition, machine learning, soft computing, computer vision, and speech recognition.

Mohammad Taghi Manzuri Shalmani received his B. Sc. and M. Sc. in Electrical Engineering from Sharif University of Technology (SUT), Iran, in 1984 and 1988, respectively. He also received the Ph. D. degree in Electrical and Computer Engineering from Vienna University of Technology, Austria, in 1995. Currently, he is an Associate Professor in Computer Engineering Department at Sharif University of Technology, Tehran, Iran. His main research interests include digital signal processing, image processing, multidimensional signal modeling, and multiresilution signal processing.