International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 4 (2012) pp. 374–382

Graph Embedding Algorithms Based on Neighborhood Discriminant Embedding for Face Recognition

Dexing Zhong[1,2], Jiuqiang Han[1], Yongli Liu[1] and Shengbin Li[2]

[1]Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University,
28 Xianning West Road, Xian, 710049 P. R. China
dxzhongstar@gmail.com

[2]State Key Laboratory of Ministry of Health for Forensic Sciences, Xian Jiaotong University,
76 Yanta West Road, Xian, 710061 P. R. China

Abstract: **This paper explores the use of a series of Graph Embedding (GE) algorithms based on Neighborhood Discriminant Embedding (NDE) as a means to improve the performance and robustness of face recognition. NDE combines GE framework and Fishers criterion and it takes into account the Individual Discriminative Factor (IDF) which is proposed to describe the microscopic discriminative property of each sample. The tensor and bilinear extending algorithms of NDE are proposed for directly utilizing the original two-dimensional image data to enhance the efficiency. The common purpose of our algorithms are to gather the within-class samples closer and separate the between-class samples further in the projected feature subspace after the dimensionality reduction. Furthermore, another informative feature extraction method called Circular Pixel Distribution (CPD) is applied to enhance the robustness of the 2-D algorithm. Experiments with the Olivetti Research Laboratory (ORL) face dataset are conducted to evaluate our methods in terms of classification accuracy, efficiency and robustness.**

Keywords: Pattern Recognition, Biometrics, Face Recognition, Graph Embedding, Neighborhood Discriminant Embedding.

I. Introduction

Face recognition has been studied since the past few decades in the field of computer vision and pattern recognition. Face patterns are subject to changes in illumination, pose and facial expression, their distribution in the high dimensional face space is nonlinear and highly complex [29]. Therefore, the way how to extract the effective discriminative features from the original image (i.e. feature extraction) is the key issue of face recognition. And many algorithms have been proposed to attempt to solve this problem. Among them, Principal Component Analysis (PCA) [1] and Linear Discriminant Analysis (LDA) [2] are two classical linear appearance-based methods. PCA is an unsupervised method which attempts to preserve the global structure of the dataset in subspace by preserving the maximum variance of data vectors. While LDA takes the class labels into consideration and searches for optimal projection directions to maximize the between-class scatter and minimize the within-class scatter at same time. Then, the kernel trick is applied to enhance the classification ability of a certain linear method (e.g. KPCA [3] and KLDA [4]). Using these methods and other derived methods [5, 6], the original 2-D face image matrices must be transformed into 1-D image vectors at first. As a result, there are two inherent limitations in these algorithms: one is that the stretching destroys the geometrical property of image space which may include some essential features for classification; the other one is that the stretched vectors lie in a high dimensional space whose value far exceeds the number of samples (high dimension but small sample size - SSS [7, 8]). Furthermore, it takes much more time to extract features from such high dimensional data which results in the considerable computational complexity.

In order to solve above problem, an algebraic feature extraction idea that directly use the image matrix to construct the image scatter for optimal discriminant analysis is proposed by Liu [9]. Inspired by this idea, Yang [10] proposes the Two-dimensional PCA (2D-PCA) to improve the performance and computational efficiency. Two-Dimensional LDA (2D-LDA) [11, 12, 13, 14] is proposed on the base of LDA. Generalized Low Rank Approximations of Matrices (GLRAM) [15, 16, 17] bears a strong resemblance to 2D-PCA in principle. The difference is GLRAM using the bilinear transportation which seeks both the left and the right projection matrices in the process of dimensionality reduction.

In 2007, a general framework [18] has been devoted to unify the dimensionality reduction algorithms over the past few decades by a general formulation known as Graph Embedding (GE). Each algorithm can be considered as the direct graph embedding or its linear/kernel/tensor extension. Two-dimensional algorithm can be considered as a 2D tensor extension. For example, Two-Dimensional Maximum Margin criterion (2D-MMC) [19] is extended from Maximum Margin criterion (MMC) [20] and Two-dimensional discriminant locality preserving projections (2D-DLPP) [21] is based on locality preserving projections (LPP) [22]. These matrix-based approaches decrease the computational complexity in the process of feature extraction and achieve better performance than classical methods.

Recently, the sparse representation of the high dimensional mass data has been introduced into the field of pattern recognition by researchers. And the classification problem is considered as a multi-class linear regression model which can be figured out by the sparse signal representation method [23]. University of Illinois Professor Y. Ma has proposed that even more discriminant information is implicit in the sparse matrix elements [24]. So with a sparse representation more important discriminant information can be extracted from this part of matrix.

In this paper, we propose a series of Graph Embedding algorithms based on Neighborhood Discriminant Embedding [25] for face recognition, which are Two-Dimensional Neighborhood Discriminant Embedding (2D-NDE), Two-Dimensional Robust Neighborhood Discriminant Embedding (2D-RNDE) and Bilinear Neighborhood Discriminant Embedding (BLNDE). 2D-NDE is a specific tensor extension of NDE which is based on graph embedding and Fisher's criterion. Individual Discriminative Factor (IDF) is introduced to give extra weightings to the critical sample points. The geometrical information of image space can be preserved during the dimensionality reduction as the original images are directly used as the input of our algorithm. Furthermore, in order to enhance the robustness of the algorithm, Circular Pixel Distribution (CPD) of face image is applied, which is a kind of description of pixel distribution of image and is considered as containing a certain two-dimensional features of the original image. And the new algorithm is called 2D-RNDE.

The rest of this paper is organized as follows: NDE and BLNDE algorithms are reviewed in Section 2 and 2D-NDE is described in Section 3. The experimental results and discussions are presented in Section 4. Finally, we conclude this study in Section 5.

II. Review of Neighborhood Discriminant Embedding (NDE) and Bilinear NDE

Given n d-dimensional samples which belong to c classes, we construct two nearest-neighbor graphs based on these n samples. They are the within-class nearest-neighbor graph G_w and the between-class nearest-neighbor graph G_b. N_i is the set consisting of k-nearest neighbor samples of the i-th sample x_i, it can be separated as the within-class set N_i^w and the between-class set N_i^b ($N_i^w \cap N_i^b = \Phi$, $N_i^w \cup N_i^b = N_i$). The Individual Discriminative Factor (IDF) is defined as follow:

$$u_i = \frac{\sum\limits_{j,x_j \in N_i^b} \exp(-\|x_i - x_j\|^2/\delta^2)}{\sum\limits_{j,x_j \in N_i} \exp(-\|x_i - x_j\|^2/\delta^2)} \tag{1}$$

where δ indicates the spread of the Gaussian.

It can be observed from this definition that: when the sample is surrounded by within-class samples and none between-class ones, the value of u_i is 0 which means the sample has almost none contribution to classifying againist other classes in a certain sense and in practical situation these samples are easy to classify and identify; when the sample is close to the border, its k-nearest neighbor samples are not all from same class, the value of u_i is nonzero. Furthermore, with the number of between-class samples increases and the distance of between-class samples decreases, the value of u_i becomes larger and the contribution to classification of these samples are greater. An extreme condition is that the value of u_i is equal to 1, which means all the k-nearest neighbor samples are from other classes and the sample is considerable far away from the within-class samples. If this kind of samples can be classified correctly that means the algorithm is very effective and robust. In summary, the value of u_i is able to represent the sample's contribution to classification in certain extent. Therefore, u_i is functional and reasonable as an individual property of each sample.

The connecting weight between two samples in G_w and G_b are s_{ij}^w and s_{ij}^b, they can be defined as follows:

$$s_{ij}^w = \begin{cases} \exp(-\|x_i - x_j\|^2/\delta^2) & x_j \in N_i^w \text{ or } x_i \in N_j^w \\ 0 & \text{others} \end{cases} \tag{2}$$

$$s_{ij}^b = \begin{cases} \exp(-\|x_i - x_j\|^2/\delta^2) & x_j \in N_i^b \text{ or } x_i \in N_j^b \\ 0 & \text{others} \end{cases} \tag{3}$$

A. Neighborhood Discriminant Embedding

In order to gather the within-class samples and separate the between-class samples after the dimensionality reduction, the following within-class and between-class objective functions are defined as:

$$\begin{aligned} J_w(W) &= \sum_{ij} \|y_i - y_j\|^2 u_i u_j s_{ij}^w \\ &= \sum_{ij} \|W^{\mathrm{T}} x_i - W^{\mathrm{T}} x_j\|^2 t_{ij}^w \\ &= tr\{W^{\mathrm{T}}(\sum_{ij}(x_i - x_j) t_{ij}^w (x_i - x_j)^{\mathrm{T}})W\} \\ &= tr\{W^{\mathrm{T}}(\begin{matrix} \sum_{ij} x_i t_{ij}^w x_i^{\mathrm{T}} + \sum_{ij} x_j t_{ij}^w x_j^{\mathrm{T}} \\ - \sum_{ij} x_i t_{ij}^w x_j^{\mathrm{T}} - \sum_{ij} x_j t_{ij}^w x_i^{\mathrm{T}} \end{matrix})W\} \\ &= 2tr\{W^{\mathrm{T}} X (D^w - T^w) X^{\mathrm{T}})W\} \\ &= 2tr\{W^{\mathrm{T}} X L^w X^{\mathrm{T}} W\} \\ &= 2tr\{W^{\mathrm{T}} S_w W\} \end{aligned} \tag{4}$$

$$J_b(W) = \sum_{ij} \|y_i - y_j\|^2 u_i u_j s_{ij}^b = 2tr\{W^{\mathrm{T}} S_b W\} \tag{5}$$

where $W \in \mathbb{R}^{d \times D}$ is the transformation matrix (d, D are the dimensions of the original sample and the reduced space, respectively) and $y_i = W^{\mathrm{T}} x_i$ is corresponding point of x_i after projection. T^w, T^b are the within-class and between-class discriminant penalty matrix , the elements of them are $t_{ij}^w = u_i u_j s_{ij}^w$, $t_{ij}^b = u_i u_j s_{ij}^b$. The symbol *"tr"* denotes the operation of trace of matrix; X is the original samples matrix, $X = [x_1, x_2, \ldots . x_n]$; $L^w = D^w - T^w, L^b = D^b - T^b$ are the within-class and between-class laplacian discriminant penalty matrices; D^w, D^b are the within-class and between-class diagonal matrices, $d_{ii}^w = \sum_j t_{ij}^w, d_{ii}^b = \sum_j t_{ij}^b$; $S_w = X L^w X^{\mathrm{T}}, S_b = X L^b X^{\mathrm{T}}$ are the within-class and between-class scatter matrices.

To minimize $J_w(W)$ and maximize $J_b(W)$ at the same time, W will be a functional and effective transformation matrix. It is equivalent to the following function:

$$W^* = \arg\max_W tr(\frac{W^{\mathrm{T}} S_b W}{W^{\mathrm{T}} S_w W}) \quad (6)$$

which is the final objective function model of neighborhood discriminant embedding (NDE).

B. Bilinear Neighborhood Discriminant Embedding

For BLNDE [26], the optimal left-multiplied and right-multiplied matrices of image matrix are derived for bilinear transformation which can preserve the geometrical property of image space and extract the row and column correlations from image matrix. After dimensionality reduction, the local within-class samples gather closer and the between-class samples separate farther. As a result, the high recognition rate is derived in the dimension-reduced feature subspace. Therefore, we define the following within-class and between-class matrix-based objective functions:

$$\min J_w(W_l, W_r) = \sum_{ij} \|Y_i - Y_j\|^2 t_{ij}^w \quad (7)$$

$$\max J_b(W_l, W_r) = \sum_{ij} \|Y_i - Y_j\|^2 t_{ij}^b \quad (8)$$

where W_l is the left-multiplied matrix and W_r is the right-multiplied matrix, $Y_i = W_l^{\mathrm{T}} X_i W_r$ (X_i is the *i*-th image sample matrix, Y_i is the projected matrix of X_i), the within-class discriminant penalty $t_{ij}^w = u_i u_j s_{ij}^w$ and the between-class discriminant penalty $t_{ij}^b = u_i u_j s_{ij}^b$. They can give the border samples extra weighting to gather the same class samples closer and separate the different class samples farther during the transformation from sample space to low-dimensional feature space.
The original image size is $h \times l$ pixels, so $X_i \in \mathbb{R}^{h\times l}$. Suppose the dimension of matrix after projection is $h_y \times l_y$, therefore $W_l \in \mathbb{R}^{h\times h_y}, W_l^{\mathrm{T}} W_l = I_{h_y}$; $W_r \in \mathbb{R}^{l\times l_y}, W_r^{\mathrm{T}} W_r = I_{l_y}$.
The basic idea of BLNDE is to find two optimal matrices W_l^* and W_r^* for Eqn. (7, 8).

$$\begin{aligned}
& \min J_w(W_l, W_r) \\
= & \sum_{ij} \|Y_i - Y_j\|^2 t_{ij}^w \quad (9)\\
= & \sum_{ij} \|W_l^{\mathrm{T}} X_i W_r - W_l^{\mathrm{T}} X_j W_r\|^2 t_{ij}^w \\
= & \sum_{ij} \|W_l^{\mathrm{T}} (X_i - X_j) W_r\|^2 t_{ij}^w \\
= & tr\{W_l^{\mathrm{T}} \sum_{ij} [(X_i - X_j) W_r t_{ij}^w W_r^{\mathrm{T}} (X_i - X_j)^{\mathrm{T}}] W_l\} \\
& \longrightarrow (a) \\
= & tr\{W_r^{\mathrm{T}} \sum_{ij} [(X_i - X_j) W_l t_{ij}^w W_l^{\mathrm{T}} (X_i - X_j)^{\mathrm{T}}] W_r\} \\
& \longrightarrow (b)
\end{aligned}$$

$$\begin{aligned}
& \max J_b(W_l, W_r) \\
= & \sum_{ij} \|Y_i - Y_j\|^2 t_{ij}^b \quad (10)\\
= & tr\{W_l^{\mathrm{T}} \sum_{ij} [(X_i - X_j) W_r t_{ij}^b W_r^{\mathrm{T}} (X_i - X_j)^{\mathrm{T}}] W_l\} \\
& \longrightarrow (a) \\
= & tr\{W_r^{\mathrm{T}} \sum_{ij} [(X_i - X_j) W_l t_{ij}^b W_l^{\mathrm{T}} (X_i - X_j)^{\mathrm{T}}] W_r\} \\
& \longrightarrow (b)
\end{aligned}$$

where the symbol *"tr"* denotes the operation of trace of matrix. Then the final objective function is presented as following:

$$\begin{aligned}
(W_l^*, W_r^*) &= \arg\max_{W_l, W_r} J(W_l, W_r) \quad (11)\\
&= \arg\max_{W_l, W_r} \{J_w(W_l, W_r)^{-1} J_b(W_l, W_r)\}
\end{aligned}$$

1) Iterative Solution Algorithm

As both $J_w(W_l, W_r)$ and $J_b(W_l, W_r)$ have two forms: (a) and (b), so the final objective function also has two forms. Generally, the Iterative Solution Algorithm is widely adopted to solve this problem. The solution can be separated as following two steps:
First Step: Given $W_r^* = (I_{l_y}, 0)^{\mathrm{T}}$ (where I_{l_y} is l_y-dimensional identity matrix and 0 is $l_y \times (l - l_y)$ zero matrix), then according to the form (a) of the final objective function, solve the optimal left-multiplied matrix W_l^*.
Second Step: As W_l^* is already known from the first step, according to the form (b) of the final objective function, solve the optimal right-multiplied matrix W_r^*.
Then, do iterative operation on above two steps, until stability results of W_l^* and W_r^* are obtained.
Observation on this algorithm, each step conducts Eigen-Decomposition on *h*-dimensional or *l*-dimensional matrix (*h*, *l* is the height and width of original image), therefore the computational complexity of each step is much lower than other vector-based approaches. Thus the times of iteration become the principal factor to the efficiency of algorithm. For this reason we give another Non-Iterative Solution Algorithm.

2) Non-Iterative Solution Algorithm

According to the experiments on the iterative solution of BLNDE, the highest recognition rate is obtained generally at 2 to 5 times. Afterwards with the increasing of loops, the recognition rate does not increase correspondingly but becomes irregular oscillation which means the optimal results are not convergent. Therefore Non-Iterative Solution Algorithm can be fulfilled and its key point is the choice of the initial condition.
In Iterative Solution Algorithm, $W_r^* = (I_{l_y}, 0)^{\mathrm{T}}$ is used as the initial condition generally. Therefore the generalized within-class scatter matrix M_w^R can be written as following:

$$\begin{aligned}
M_w^R &= \sum_{ij} (X_i - X_j) W_r^* t_{ij}^w W_r^{*\mathrm{T}} (X_i - X_j)^{\mathrm{T}} \\
&= \sum_{ij} \overline{X}_{ij} (I_{l_y}, 0)^{\mathrm{T}} t_{ij}^w (I_{l_y}, 0) \overline{X}_{ij}^{\mathrm{T}}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{ij} \left[\overline{X}_{ij}^1|_{h\times l_y}, \overline{X}_{ij}^2|_{h\times(l-l_y)}\right]_{h\times l} \\
&\quad \begin{bmatrix} I_{l_y\times l_y} \\ \Phi_{(l-l_y)\times l_y} \end{bmatrix} t_{ij}^w \left[I_{l_y\times l_y}, \Phi_{l_y\times(l-l_y)}\right] \\
&\quad \begin{bmatrix} (\overline{X}_{ij}^1)^{\mathrm{T}}|_{l_y\times h} \\ (\overline{X}_{ij}^2)^{\mathrm{T}}|_{(l-l_y)\times h} \end{bmatrix}_{l\times h} \\
&= \sum_{ij} \overline{X}_{ij}^1|_{h\times l_y} I_{l_y\times l_y} t_{ij}^w I_{l_y\times l_y} (\overline{X}_{ij}^1)^{\mathrm{T}}|_{l_y\times h} \\
&= \sum_{ij} t_{ij}^w \overline{X}_{ij}^1|_{h\times l_y} \cdot (\overline{X}_{ij}^1)^{\mathrm{T}}|_{l_y\times h} \qquad (12)
\end{aligned}
$$

where $\overline{X}_{ij} = X_i - X_j = \left[\overline{X}_{ij}^1|_{h\times l_y}, \overline{X}_{ij}^2|_{h\times(l-l_y)}\right]_{h\times l}$.
According to Eqn. (12), M_w^R is just relevant to the first part of image matrix $\overline{X}_{ij}$. As a result, the input of original image information is incomplete. The rest part information is added in later through iteration. Therefore it is not the most efficient approach.

In this paper, we break the dimensional limitation of the initial value of $W_r \in \mathbb{R}^{l\times l_y}$, and offer an expanded matrix for the initialization of W_r^*.

$$
W_r^* = I_{l\times l} = \begin{bmatrix} I_{l_y\times l_y} & \Phi_{l_y\times(l-l_y)} \\ \Phi_{(l-l_y)\times l_y} & I_{(l-l_y)\times(l-l_y)} \end{bmatrix}_{l\times l} \qquad (13)
$$

Consequently, M_w^R is written as following:

$$
\begin{aligned}
M_w^R &= \sum_{ij} (X_i - X_j) W_r^* t_{ij}^w W_r^{*\mathrm{T}} (X_i - X_j)^{\mathrm{T}} \\
&= \sum_{ij} \overline{X}_{ij} I_{l\times l} t_{ij}^w I_{l\times l}^{\mathrm{T}} \overline{X}_{ij}^{\mathrm{T}} \\
&= \sum_{ij} \left[\overline{X}_{ij}^1|_{h\times l_y}, \overline{X}_{ij}^2|_{h\times(l-l_y)}\right]_{h\times l} \\
&\quad \begin{bmatrix} I_{l_y\times l_y} & \Phi_{l_y\times(l-l_y)} \\ \Phi_{(l-l_y)\times l_y} & I_{(l-l_y)\times(l-l_y)} \end{bmatrix}_{l\times l} t_{ij}^w \\
&\quad \begin{bmatrix} I_{l_y\times l_y} & \Phi_{l_y\times(l-l_y)} \\ \Phi_{(l-l_y)\times l_y} & I_{(l-l_y)\times(l-l_y)} \end{bmatrix}_{l\times l} \\
&\quad \begin{bmatrix} (\overline{X}_{ij}^1)^{\mathrm{T}}|_{l_y\times h} \\ (\overline{X}_{ij}^2)^{\mathrm{T}}|_{(l-l_y)\times h} \end{bmatrix}_{l\times h} \\
&= \sum_{ij} \left[\overline{X}_{ij}^1|_{h\times l_y}, \overline{X}_{ij}^2|_{h\times(l-l_y)}\right]_{h\times l} t_{ij}^w \\
&\quad \begin{bmatrix} (\overline{X}_{ij}^1)^{\mathrm{T}}|_{l_y\times h} \\ (\overline{X}_{ij}^2)^{\mathrm{T}}|_{(l-l_y)\times h} \end{bmatrix}_{l\times h} \\
&= \sum_{ij} t_{ij}^w \overline{X}_{ij}|_{h\times l} \cdot \overline{X}_{ij}^{\mathrm{T}}|_{l\times h} \qquad (14)
\end{aligned}
$$

According to Eqn. (14), using the expanded matrix as the initial value of W_r^*, shown in Eqn. (13), all the elements of the original image matrix are involved which means the one-time input of overall image information at the initiation. According to the above improved initiation, the results of one-time iteration are used as the final optimal W_l^* and W_r^*. Thus the computational complexity is significantly lowered. In addition, the experiments result in [26] demonstrate that the image features extracted by Non-Iterative Solution Algorithm have more effective ability to classification than the ones of Iterative Solution.

III. Two-Dimensional Robust Neighborhood Discriminant Embedding

A. 2D Neighborhood Discriminant Embedding

NDE stretches the orignal image data into a vector at the beginning. In this situation the geometrical property of image space does not be taken into consideration, therefore some crucial information maybe lost during the transformation. In order to avoid the loss, Two-dimensional Neighborhood Discriminant Embedding (2D-NDE) is proposed as an image matrix-based algorithm which utilizes the similar idea of NDE.

Given n image matrix training samples and they belong to c classes. The original image size is $h \times l$ pixels. X_i is the i-th samples, $X_i \in \mathbb{R}^{h\times l}$. The projection is fulfilled through the right-multiplied transformation matrix W_r, $W_r \in \mathbb{R}^{l\times l_y}$, where l_y is the column number of the extracted image feature matrix after the projection, and $Y_i = X_i W_r \in \mathbb{R}^{h\times l_y}$.

First, we give the expanded within-class objective function as follows:

$$
\begin{aligned}
J_w(W_r) &= \sum_{ij} \|Y_i - Y_j\|^2 t_{ij}^w \qquad (15) \\
&= \sum_{ij} \|X_i W_r - X_j W_r\|^2 t_{ij}^w \\
&= \sum_{ij} \|(X_i - X_j) W_r\|^2 t_{ij}^w \\
&= tr\{W_r^{\mathrm{T}} [\sum_{ij} (X_i - X_j)^{\mathrm{T}} t_{ij}^w (X_i - X_j)] W_r\}
\end{aligned}
$$

where t_{ij}^w is the within-class connected weight defined in NDE, $t_{ij}^w = u_i u_j s_{ij}^w$.

Here in order to facilitate the solution, we define the expanded within-class discriminant penalty matrix T_E^w, the expanded within-class diagonal matrix D_E^w and the expanded within-class laplacian discriminant penalty matrix L_E^w. $T_E^w, D_E^w, L_E^w \in \mathbb{R}^{hn\times hn}$.

$$
T_E^w = \begin{bmatrix} T_{11}^w & \cdots & T_{1n}^w \\ \vdots & \ddots & \vdots \\ T_{n1}^w & \cdots & T_{nn}^w \end{bmatrix} = \begin{bmatrix} t_{11}^w \cdot I_h & \cdots & t_{1n}^w \cdot I_h \\ \vdots & \ddots & \vdots \\ t_{n1}^w \cdot I_h & \cdots & t_{nn}^w \cdot I_h \end{bmatrix} \qquad (16)
$$

$$
D_E^w = \begin{bmatrix} D_{11}^w & & 0 \\ & \ddots & \\ 0 & & D_{nn}^w \end{bmatrix} \qquad (17)
$$

$$
L_E^w = D_E^w - T_E^w \qquad (18)
$$

where $D_{ii}^w = d_{ii}^w \cdot I_h, d_{ii}^w = \sum_j t_{ij}^w$ which is the element of D^w defined in NDE.

Then equation (15) can be written as:

$$
\begin{aligned}
J_w(W_r) &= \sum_{ij} \|Y_i - Y_j\|^2 t_{ij}^w \qquad (19) \\
&= tr\{W_r^{\mathrm{T}} (2X_E D_E^w X_E^{\mathrm{T}} - 2X_E T_E^w X_E^{\mathrm{T}}) W_r\} \\
&= 2tr\{W_r^{\mathrm{T}} (X_E (D_E^w - T_E^w) X_E^{\mathrm{T}}) W_r\} \\
&= 2tr\{W_r^{\mathrm{T}} (X_E L_E^w X_E^{\mathrm{T}}) W_r\} \\
&= 2tr\{W_r^{\mathrm{T}} S_w^I W_r\}
\end{aligned}
$$

where $X_E = [X_1^{\mathrm{T}}, X_2^{\mathrm{T}}, \ldots, X_n^{\mathrm{T}}]$, $S_w^I = X_E L_E^w X_E^{\mathrm{T}}$ which is the within-class scatter matrix.
Likewise, we can define the expanded between-class objective function as:

$$\begin{aligned} J_b(W_r) &= \sum_{ij} \|Y_i - Y_j\|^2 t_{ij}^b \\ &= \sum_{ij} \|X_i W_r - X_j W_r\|^2 t_{ij}^b \end{aligned} \tag{20}$$

where $t_{ij}^b = u_i u_j s_{ij}^b$, and we define the expanded matrices: $T_E^b, D_E^b, L_E^b \in \mathbb{R}^{hn \times hn}$.

$$T_E^b = \begin{bmatrix} T_{11}^b & \cdots & T_{1n}^b \\ \vdots & \ddots & \vdots \\ T_{n1}^b & \cdots & T_{nn}^b \end{bmatrix} = \begin{bmatrix} t_{11}^b \cdot I_h & \cdots & t_{1n}^b \cdot I_h \\ \vdots & \ddots & \vdots \\ t_{n1}^b \cdot I_h & \cdots & t_{nn}^b \cdot I_h \end{bmatrix} \tag{21}$$

$$D_E^b = \begin{bmatrix} D_{11}^b & & 0 \\ & \ddots & \\ 0 & & D_{nn}^b \end{bmatrix} \tag{22}$$

$$L_E^b = D_E^b - T_E^b \tag{23}$$

and equation (19) can be written as:

$$\begin{aligned} J_b(W_r) &= \sum_{ij} \|Y_i - Y_j\|^2 t_{ij}^b \\ &= tr\{W_r^{\mathrm{T}}(2X_E D_E^b X_E^{\mathrm{T}} - 2X_E T_E^b X_E^{\mathrm{T}})W_r\} \\ &= 2tr\{W_r^{\mathrm{T}}(X_E L_E^b X_E^{\mathrm{T}})W_r\} \\ &= 2tr\{W_r^{\mathrm{T}} S_b^I W_r\} \end{aligned} \tag{24}$$

where $S_b^I = X_E L_E^b X_E^{\mathrm{T}}$ which is the between-class scatter matrix.
Hereby, the within-class and between-class scatter matrices have the same form with the ones in NDE which bases on the image vector. The final objective function model of two-dimensional neighborhood discriminant embedding (2D-NDE) is as follows:

$$\begin{aligned} W_r^* &= \arg\max_{W_r} J(W_r) \\ &= \arg\max_{W_r} tr \frac{W_r^{\mathrm{T}} S_b^I W_r}{W_r^{\mathrm{T}} S_w^I W_r} \end{aligned} \tag{25}$$

The solution of equation (25) is equivalent to the solution of Generalized Characteristic Equation [27]:

$$S_b^I w_r = \lambda S_w^I w_r \tag{26}$$

where w_r is the column vector of W_r.
Usually S_w^I is a nonsingular matrix, we can work out directly by the Eigen-Decomposition of $(S_w^I)^{-1} S_b^I$. We use the eigenvectors from the first to the l_y-th largest eigenvalue to compose a matrix, which is the optimal right-multiplied transformation matrix W_r^*. As the dimension of S_w^I and S_b^I are much lower than S_w and S_b constructed by vector-based methods, the computational complexity is decreased significantly.

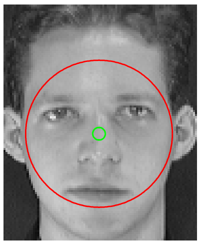

Figure. 1: Here the process of CPD characterization is shown. The red large circle is the smallest circle which is able to cover eyebrows, eyes, noise and mouth. The integer R is an approximation of the radius of the smallest circle. The green small circle presents the centre. We sum the pixel value of the points of each circular to construct a R-dimensional vector.

B. Two-Dimensional Robust Neighborhood Discriminant Embedding (2D-RNDE)

Observed in the later experiments, 2D-NDE can achieve a considerable and comparable recognition accuracy at certain dimensional feature subspace, but cannot keep this accuracy across different dimensionality of the projected subspace. One of the reason maybe that 2D-NDE just considers the column correlation of the original image data. So that more geometrical property of the original image should be involved. Inspired by Circular Pixel Distribution (CPD), an improved algorithm, named as Two-Dimensional Robust Neighborhood Discriminant Embedding (2D-RNDE), has been proposed to enhance the robustness of 2D-NDE.
Every face image being analyzed has its own Circular Pixel Distribution which contains another kind of geometrical property and can be directly used as an additional feature. Fig. 1 shows the process of calculating CPD of face image. Circles were drawn around a certain centre of the face image, which is also the centre of the smallest circle that is able to cover eyebrows, eyes, noise and mouth. And an integer R is an approximation of the radius of the smallest circle. Around the centre we can build R circulars with the radius increasing one by one. We sum the pixel value of the points of each circular to construct a R-dimensional vector. In this way, all the pixels inside the smallest circle which covers main part of the face are involved into the vector construction. And this vector is directly applied as a supplement to the features extracted by 2D-NDE, then the new feature is used for classification. Combining 2D-NDE and CPD, 2D-RNDE can achieve better performance of robustness.

IV. The experimental results and discussions

A. Compare 2D-NDE, 2D-RNDE with 2D-PCA

In this section, we compare our proposed methods with 2D-PCA on ORL [28] database. The ORL (Olivetti Research Laboratory) face database consists of 400 face images (total

Figure. 2: Example of the ORL database, 2 persons and each has 10 images.

40 people and 10 samples per person). The images contain a certain amount change of light, different facial expression, facial details (glasses or no glasses), as well as a range of in-depth rotation. An example of ORL is shown in Fig. 2. For each individual, $\xi(=3,4,5)$ images are randomly selected for training and the rest are used for testing. Nearest Neighbor Classifier is applied in this paper for classification. In order to reduce the random error and give a more objective evaluation of the performance of the algorithms, for each ξ we do 50 times random selection of the training set of samples and take the average accuracy as the final results. The radius of the smallest circle $R = 35$ in our experiments.

Fig. 3 illustrates the comparison among 2D-PCA, 2D-NDE and 2D-RNDE on ORL in three cases ($\xi = 3,4,5$). Then the highest recognition rate and the corresponding number of image feature's column of them in each case are made into a table, shown in Table 1.

Table 1: Comparison of top recognition rate of 2D-PCA, 2D-NDE and 2D-RNDE on ORL

Method	$\xi = 3$	$\xi = 4$	$\xi = 5$
2D-PCA	88.57%(6)	91.67%(6)	93.00%(7)
2D-NDE	92.14%(3)	93.75%(5)	94.00%(8)
2D-RNDE	92.86%(8)	93.75%(5)	94.00%(9)

In Fig. 3, it is very obvious that our proposed 2D-NDE and 2D-RNDE outperform 2D-PCA almost across all the values of dimensionality of the reduced space. 2D-PCA only uses the global geometric information of Euclidean space besides the column correlation. While 2D-NDE and 2D-RNDE inherit directly from NDE which is derived as the projected directions that preserve the local neighborhood structure on the data manifold of the face space. Meanwhile, benefiting from Graph Embedding and Fisher's criterion, NDE can gather the within-class samples closer and separate the between-class samples farther in the projected space. As a specific 2D tensor extension of NDE, 2D-NDE takes into account the column correlation information and utilizes the image matrix directly, as a result the computational complexity and time-consuming are decreased and the classification accuracy is kept at the considerable level. However, 2D-NDE can achieve the high recognition accuracy at certain dimensional feature subspace, but cannot keep this accuracy across different dimensionality of projected subspace. Utilizing CPD, another kind of geometric property of face image is involved in our algorithm, 2D-RNDE can get the same accuracy as 2D-NDE but better robustness across the dimensionality of subspace.

B. Comparison among the NDE algorithms

We conduct a comprehensive comparison among NDE, BLNDE, 2D-NDE and 2D-RNDE on ORL face database, and compare NDE algorithms with other face recognition methods, based on these comparisons we can draw an objective evaluation on NDE algorithms.

Table 2: Comparison of top recognition rate of NDE algorithms on ORL

Method	$\xi = 3$	$\xi = 4$	$\xi = 5$
NDE	91.5%(83)	**94.6%(45)**	**96.2%(56)**
2D-NDE	92.14%(3)	93.75%(5)	94.00%(8)
2D-RNDE	**92.86%(8)**	93.75%(5)	94.00%(9)
Iterative BLNDE	86.79%(13,10)	91.92%(13,10)	94.50%(13,10)
Non-Iter. BLNDE	92.03%(12,8)	94.08%(12,8)	96.00%(12,8)

Note: Values in parentheses are the corresponding feature dimension that the highest accuracy achieved, the maximum value of each column is in bold.

NDE and other four derived algorithms are compared on the ORL face database in three cases of the training sample set, shown in Table 2. In the case of insufficient training samples ($\xi = 3$), three samples of each class are randomly selected as the training sample set, 2D-RNDE achieved the highest correct recognition rate (92.86%). With the increase in the number of training samples ($\xi = 4,5$), i.e. randomly selecting 4 or 5 samples from each class as the training sample set, NDE achieved the highest recognition rate, which are 94.6% and 96.2%. With the increase in the number of training samples, the accuracies of NDE algorithms have a corresponding increase, which is also consistent with the intuitive idea that the greater number of training samples can enhance the algorithm's ability to classification.

In the case of insufficient training samples, the introduction of the column or bilinear correlation of face image matrix can enhance the correct recognition rate, therefore 2D-NDE, 2D-RNDE and Non-Iterative BLNDE have achieved better results than NDE. The introduction of CPD is even more helpful to improve the performance of 2D-NDE, including the accuracy and robustness. This shows that in case of the inadequate training samples, the method based on image matrix is more effective than image vector-based approach.

With the increase in the number of training samples, NDE the method based on image vector can construct more complete high-dimensional manifold data structure of the local

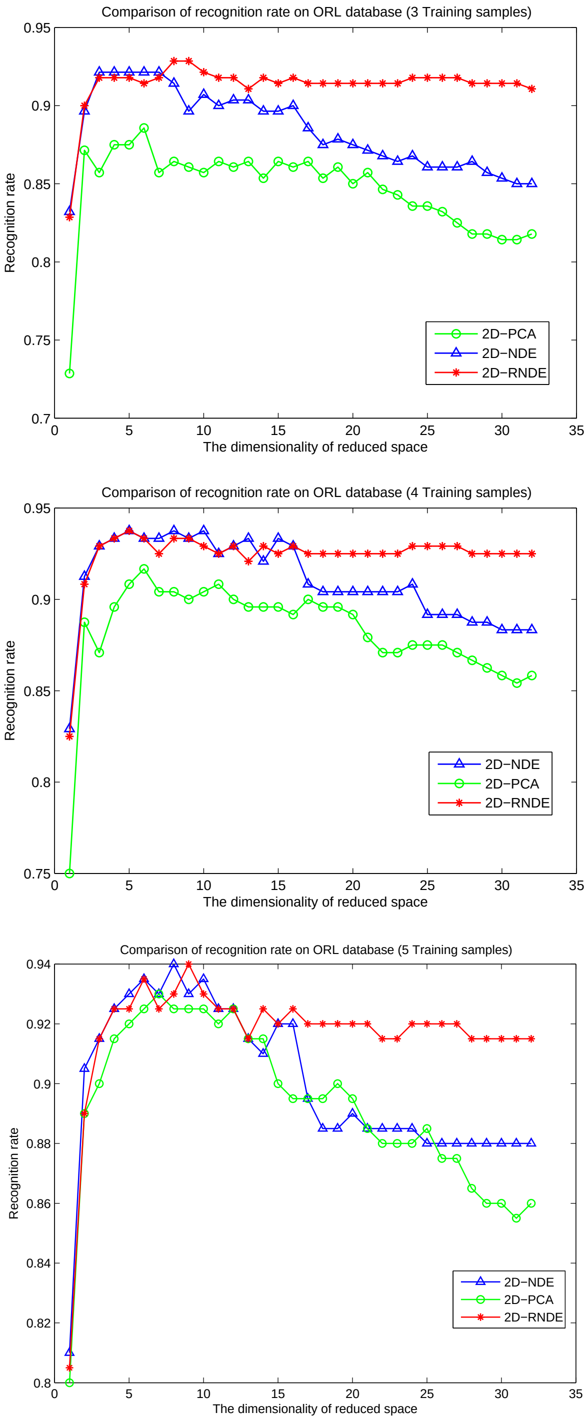

Figure. 3: Comparison among 2D-PCA, 2D-NDE and 2D-RNDE on ORL

neighborhood, while the image matrix-based approaches do not stretch the original image data to a high-dimensional vector, so they are not fully benefited from the more complete of manifold structure. Therefore the accuracy of these approaches are not better than the image vector-based method. However, to compare the efficiency of algorithm's implementation, the advantage of image matrix-based methods will be very obvious. For example, each sample of ORL is 112×92 face image, the vector for NDE to process will reach 10^4, and if we do 50 times randomly sample selecting experiment, the time consumption is calculated measured with hours. Meanwhile, the image matrix-based approaches can fulfill the same calculations in a few minutes. Specifically, the average time consumption of each algorithm is shown in Table 3.

Table 3: Comparison of average time of 2D-NDE, 2D-RNDE and BLNDE (Unit: seconds)

Method	2D-NDE	2D-RNDE	It. BLNDE	Non-It. BLNDE
Time	6.87	8.36	12.52	7.07

C. *Compare the NDE algorithms with other 6 methods*

We compare NDE algorithms with other face recognition approaches. In ORL face database, using five randomly selected samples of each class as the training sample set is considered as the compromise strategy between computing time and recognition rate, so we also used the same way to execute the following experiments. Comparison of NDE algorithms and six other algorithms is shown in Table 4.

Table 4: Comparison of average time of 2D-NDE, 2D-RNDE and BLNDE (Unit: seconds)

PCA	LDA	LPP	NDE	
91.5%(49)	92.2%(39)	95.5%(46)	**96.2%(56)**	
2D-PCA	2D-LDA	2D-LPP	2D-NDE	BLNDE
93.0%(7)	94.0%(12,8)	95.5%(40)	94.0%(8)	96.0%(12,8)

Note: Percentages express the highest recognition rate of algorithms, and the value in parentheses represent the corresponding feature dimension, in which the two values mean the corresponding row and column of feature image. The maximum is in bold. BLNDE takes the Non-Iterative algorithm.

Actually, the nine methods in Table 4 can be classified according to different principles.

1. Classification according to prototype
   PCA, LDA, 2D-PCA and 2D-LDA are based on Eigenfaces; LPP and 2D-LPP are based on LPP; NDE, 2D-NDE and BLNDE are based on NDE.

2. Classification according to 1D or 2D method
   PCA, LDA, LPP and NDE are 1D methods based on image vector; 2D-PCA, 2D-LDA, 2D-LPP, 2D-NDE and BLNDE are 2D methods based on image matrix, in which 2D-LDA and BLNDE are bilinear methods.

3. Classification according to using Discriminant Analysis or not
   PCA, LPP, 2D-PCA and 2D-LPP are without discriminant analysis, i.e. the algorithms do not use the class label of samples; LDA, NDE, 2D-LDA, 2D-NDE and BLNDE take use of the class label of samples and so that they are with discriminant analysis.

The nine methods can be compared according to the above three categories. For comparison, we use the inequality to compare two algorithms, for example, $a \geq b$ means the accuracy of Algorithm a is greater than or equal to Algorithm b. Comparison results are as follows:

1. $NDE \geq LPP \geq LDA \geq PCA; BLNDE \geq 2D - LPP \geq 2D - LDA \geq 2D - PCA$

2. $2D - PCA \geq PCA; 2D - LDA \geq LDA; 2D - LPP \geq LPP; NDE \geq 2D - NDE$
3. $NDE \geq LPP; LDA \geq PCA; BLNDE \geq 2D - LPP; 2D - LDA \geq 2D - PCA$

The above comparison can be summarized as:

1. Algorithm based on NDE $\geq$ Algorithm based on LPP $\geq$ Algorithm based on Eigenfaces
2. 2D methods $\geq$ 1D methods, except NDE
3. Algorithm with label $\geq$ Algorithm without label

Eigenfaces methods can retain the global structure of the training samples; LPP methods are to find the local geometric properties of the training samples; NDE methods combine graph embedding framework and the Fisher criterion to deal with within-class and between-class discriminative information separately. In the face recognition field, the local geometric properties may be more influential than the global structure, and the distinction within-class and between-class discriminant information of the local geometry can further improve the algorithm's discriminative ability.

1D methods need to stretch the original face image matrix into vector, which often leads to high dimensional vector space, and the geometric properties of face image will be destroyed, a lot of information about the spatial structure will lost in the process as well. However those information may play a more important role in the search of the most optimal projection direction. By comparison, 2D methods can effectively avoid these shortcomings, thereby enhancing the ability of classification. NDE does not fully meet this principle, only in the case that the training sample is not sufficient $(\xi = 3), 2D - NDE \geq NDE$.

Using discriminant analysis with class label is very effective for recognition problem, because the discriminant analysis methods are based on known class labels of training samples to construct the projection model, this approach can maximize the distance of between-class samples and minimize the distance of within-class samples, thereby enhancing the ability of the discriminant algorithm.

According to the above analysis, the reason that higher recognition rate can be obtained by NDE algorithms is simple and intuitive. NDE methods combine graph embedding and Fisher criterion, take full use of within-class and between-class discriminant information of the local geometry, and consider the class labels using discriminant analysis.

V. Conclusions

In this paper, 2D-NDE was developed based on three techniques, i.e. Neighborhood Discriminant Embedding (NDE), matrix-based projection and discriminant analysis. Experimental results and comprehensive comparisons demonstrated that the above three techniques play active roles in enhancing the performances of algorithms for face recognition. Firstly, NDE is the foundation of 2D-NDE, which incorporates Graph Embedding and Fisher's Criterion as well as including an Individual Discriminative Factor (IDF) as an individual property of each sample to describe the contribution to discriminant. Moreover, NDE is more effective than LPP and Eigenfaces in searching the optimal projection direction. Secondly, matrix-based projection does not need to stretch the original face image matrix into vector, so that it can avoid the destroying of the geometric properties of image and small sample size (SSS) problem. Thirdly, discriminant analysis is utilized to deal with the within-class and between-class discriminant information of local geometry separately, which can further improve the algorithm's discriminative ability. Additionally, 2D-RNDE is proposed by employing Circular Pixel Distribution (CPD) to include extra geometric property of face image, and it leads to the enhancement of the robustness of 2D-NDE. The local information of underlying manifold structure, the between-class scatter and the within-class scatter, the column correlations of image matrix and circular pixel distribution are taken into consideration all-in-one. As a result, face recognition can be better conducted in the projected feature subspace after the dimensionality reduction. Experimental results on ORL face database demonstrate the effectiveness and robustness of our algorithms.

Acknowledgments

This work was supported by grants from National Natural Science Foundation of China (NSFC, No. 61105021), China Postdoctoral Science Foundation (No. 2011M501442) and the Fundamental Research Funds for the Central Universities.

References

[1] M. Turk, and A. Pentland, Eigenfaces for recognition, *Cognitive Neuroscience*, 3, pp. 71-86, 1991.

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, *IEEE Trans. PAMI*, 19 (7), pp. 711-720, 1997.

[3] B. Scholkopf, A. Smola, and K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10, pp. 1299-1319, 1998.

[4] S. Mika, G. Ratsch, and J. Weston, Fisher discriminant analysis with kernels. In *Proc. IEEE Neural Networks for Signal Processing*, pp. 41-48, 1999.

[5] X. F. He, S. C. Yan, Y. X. Hu, Face recognition using laplacian faces, *IEEE. Trans. Pattern Analysis and Machine Intelligence*, 27 (3), pp. 328-340, 2005.

[6] Q. B. You, N. N. Zheng, S. Y. Du, Neighborhood discriminant projection for face recognition, *Pattern Recognition Letters*, 28 (10), pp. 1156-1163, 2007.

[7] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data, *Appl. Stat.*, 44 (1), pp. 101-115, 1995.

[8] S. J. Raudys, and A. K. Jain, Small sample size effects in statistical pattern recognition: Recommendations for practitioners, *IEEE Trans. Pattern Anal. Mach. Intell.*, 13 (3), pp. 252-264, 1991.

[9] K. Liu, Y. Q. Cheng, and J. Y. Yang, Algebraic feature extraction for image recognition based on an optimal discriminant criterion, *Pattern Recognition*, 26 (6), pp. 903-911, 1993.

[10] J. Yang, D. Zhang, A. F. Frangi, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26 (1), pp. 131-137, 2004.

[11] M. Li, and B. Yuan, 2D-LDA: A statistical linear discriminant analysis for image matrix, *Pattern Recognition Letters*, 26, pp. 527-532, 2005.

[12] J. P. Ye, R. Janardan, and Q. Li, Two-dimensional linear discriminant analysis. In *The Eighteenth Annual Conference on Neural Information Processing Systems*, pp. 1569-1576, 2004.

[13] J. Yang, D. Zhang, X. Yong, Two-dimensional discriminant transform for face recognition, *Pattern Recognition*, 38 (7), pp. 1125-1129, 2005.

[14] S. Noushath, G. H. Kumar, and P. Shivakumara, (2D)(2) LDA: An efficient approach for face recognition, *Pattern Recognition*, 39 (7), pp. 1396-1400, 2006.

[15] J. P. Ye, Generalized low rank approximations of matrices, *Machine Learning*, 61 (1-3), pp. 167-191, 2005.

[16] Z. Z. Liang, D. Zhang, and P. F. Shi, The theoretical analysis of GLRAM and its applications, *Pattern Recognition*, 40 (3), pp. 1032-1041, 2007.

[17] C. Lu, W. Q. Liu, and S. J. An, A simplified GLRAM algorithm for face recognition, *Neurocomputing*, 72 (1-3), pp. 212-217, 2008.

[18] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, and S. Lin, Graph embedding and extensions: A general framework for dimensionality reduction, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, pp. 40-51, 2007.

[19] Wen-Hui Yang, and Dao-Qing Dai, Two-Dimensional Maximum Margin Feature Extraction for Face Recognition, *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, 39 (4), 1002-12, 2009.

[20] H. Li, T. Jiang, and K. Zhang, Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Netw.*, 17 (1), pp. 157-165, 2006.

[21] Yu Weiwei, Two-dimensional discriminant locality preserving projections for face recognition, *Pattern Recognition Letters*, 30, pp. 1378-1383, 2009.

[22] H. Xiaofei, Y. Shuicheng, H. Yuxiao, P. Niyogi, and Z. Hong-Jiang, Face recognition using Laplacianfaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, pp. 328-40, 2005.

[23] R. Zhi, M. Flierl, Q. Ruan et al., Graph-Preserving Sparse Nonnegative Matrix Factorization With Application to Facial Expression Recognition, *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, 41 (1), pp. 38-52, 2011.

[24] E. J. Candes, X. Li, Y. Ma et al., Robust Principal Component Analysis?, *Journal of the ACM*, 58 (3), pp. 11, 2011.

[25] D. Zhong, J. Han, X. Zhang, and Y. Liu, Neighborhood discriminant embedding in face recognition, *Optical Engineering*, 49 (7), 077203, 2010.

[26] D. Zhong, J. Han, and Y. Liu, Bilinear Neighborhood Discriminant Embedding and its Non-Iterative Solution Algorithm. In *Proc. of the 2nd International Congress on Image and Signal Processing (CISP'09)*, 2009.

[27] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7, pp. 179-188, 1936.

[28] F.S. Samaria, and A.C. Harter, Parameterisation of a stochastic model for human face identification. In *Proc. Second IEEE Workshop on Applications of Computer Vision*, pp. 138-142, 1994.

[29] K. R. Singh, M. A. Zaveri, M. M. Raghuwanshi, Illumination and Pose Invariant Face Recognition: A Technical Review, *International Journal of Computer Information Systems and Industrial Management*, 2, pp. 29-38, 2010.

Author Biographies

Dexing Zhong was born in TianChang, Anhui Province of China in 1983. He received his B.S. degree in information systems and Ph.D. degree in control science and engineering from Xi'an Jiaotong University, China in 2005 and 2010, respectively. He was a visiting scholar with the University of Illinois, Urbana-Champaign, from 2007 to 2008. He is a lecturer in the School of Electronic and Information Engineering, Xi'an Jiaotong University. His research interests include pattern recognition, computer vision, bioinformatics and sensor networks.

Jiuqiang Han graduated from Xi'an Jiaotong University, China, where he joined the faculty in 1977 and is currently a professor in the School of Electronic and Information Engineering. His current research interests include 3-D image measurement and fusion, computer vision, pattern recognition, network measurement and control, and sensor networks.

Yongli Liu received her BS and MS degrees in information systems and control science and engineering from Xian Jiaotong University in 2005 and 2008, respectively. Her research interests include face recognition, face cartoons, and facial expression transformation.