A Novel Approach to Expression Recognition from Non-frontal Face Images

Wenming Zheng[1,2], Hao Tang[1], Zhouchen Lin[3], Thomas S. Huang[1]
[1]Beckman Institute, University of Illinois at Urbana-Champaign, USA
[2]Research Center for Learning Science, Southeast University, Nanjing 210096, China.
[3]Vision Computing Group, Microsoft Research Asia, China
E-mail: wenming_zheng@seu.edu.cn

Abstract

Non-frontal view facial expression recognition is important in many scenarios where the frontal view face images may not be available. However, few work on this issue has been done in the past several years because of its technical challenges and the lack of appropriate databases. Recently, a 3D facial expression database (BU-3DFE database) is collected by Yin et al. [10] and has attracted some researchers to study this issue. Based on the BU-3DFE database, in this paper we propose a novel approach to expression recognition from non-frontal view facial images. The novelty of the proposed method lies in recognizing the multi-view expressions under the unified Bayes theoretical framework, where the recognition problem can be formulated as an optimization problem of minimizing an upper bound of Bayes error. We also propose a close-form solution method based on the power iteration approach and rank-one update (ROU) technique to find the optimal solutions of the proposed method. Extensive experiments on BU-3DFE database with 100 subjects and 5 yaw rotation view angles demonstrate the effectiveness of our method.

1. Introduction

Automatic facial expression recognition has become a very hot research topic in computer vision and pattern recognition community due to its technical challenge and the wide potential applications in many fields. A major task of facial expression recognition is to classify a given facial image into six categories, i.e. angry, disgust, fear, happy, sad, and surprise, based on some facial features. During the past several years, many facial expression recognition methods have been proposed. For a literature survey, see [3][8][11]. A major limitation of the previous facial expression recognition methods is that most of them focus on the frontal or nearly frontal view facial images. In many scenarios, however, frontal or nearly frontal view face images may not be available. Moreover, people have found that non-frontal facial images may be more informative than frontal ones because in non-frontal facial images the heights of the nose and cheek may be available, while frontal ones are usually symmetric and redundant.

Recently, a publicly available 3D facial expression database, called BU-3DFE, was collected by Yin et al. [10] at State University of New York at Binghamton. The BU-3DFE facial expression database attracted some researchers to investigate the non-frontal view facial expression recognition issue and had obtained some interesting findings. For example, Wang et al. [9] found that the expression recognition performed poorly when the facial images suffered from large view variation by comparing the expression recognition performance under the different testing facial view regions based on the classifier trained on the frontal-view facial images. Another interesting finding was conducted by Hu et al. [6] who conducted various facial expression experiments on five yaw views, i.e., 0^o, 30^o, 45^o, 60^o, and 90^o. They found that using the face images under the non-frontal view can obtain better recognition performance than under the frontal view. The major limitation of their non-frontal view approache is that they are view-dependent between the training and testing face images, i.e., they train the algorithm with the face images of some specific views. Hence, they need to know the viewing angles of the testing images before the recognition procedure is performed, which is still a challenging work in computer vision.

In this paper, we propose a novel approach to the expression recognition from non-frontal view facial images. The novelty of the proposed method lies in recognizing the multi-view expressions under the unified Bayes theoretical framework. More specifically, we formulate the multi-view facial expression recognition problem as the optimization problem of minimizing an upper bound of Bayes error, which boils down to solve a series of principal eigenvector of matrices, which can be efficiently solved via the power iteration approach [5] and rank-one update (ROU) technique. Moreover, to obtain better recognition performance, we use the SIFT (Scale Invariant Feature Transform) features [7] to represent the face images. The SIFT features are known to be invariant to image scale and rotation, and also robust across changes in illumination, noise, and a substantial range of affine distortion [7]. In this paper, we extract the sparse SIFT features loacted 83 facial feature points (FPs) of each face image to represent the the face images.

The rest of this paper is organized as follows: In section 2, we address the BU-3DFE database and the SIFT features representation. In section 3, we address our non-frontal view facial expression recognition approach. The experiments are presented in section 4. Finally section 5 concludes our paper.

2. Description of the BU-3DFE Database and SIFT Features Representation

The BU-3DFE database consists of 100 subjects (56 female and 44 male) of different ethnicities, each of whom elicits 6 universal facial expressions (anger, disgust, fear, happiness, sadness, and surprise) with 4 levels of intensities. The 2400 3D facial expression models are described by both 3D geometrical shapes and color textures. To facilitate correspondences, 83 FPs are identified on every 3D model. A more detailed description of the database can be found in [10]. We render the 3D models with OpenGL by selecting proper viewpoints, resulting in 5-view projected face images corresponding to 0^o, 30^o, 45^o, 60^o, and 90^o yaw angles. As a result, the 83 FPs are also projected onto the same 2D faces. Fig. 1 shows some examples of the 2D face images rendered from the 3D face models.

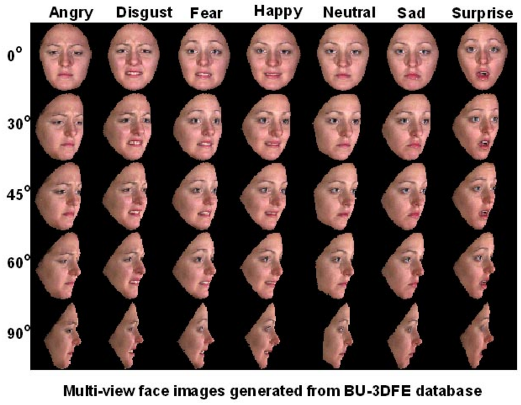

Figure 1. Some facial images of different expressions obtained by projecting 3D face models in BU-3DFE onto different views.

For each 2D face image, we extract the sparse SIFT features located in 83 FPs to represent that image, where the sparse SIFT features are extracted as follows:

1. For each 2D face image, we use the 83 FPs as the key points where the SIFT features to be computed. If the key points are occluded by the face, we simply compute the SIFT features in the corresponding positions of the image.
2. In computing the SIFT features, the same fixed horizontal orientation is used for all the 83 key points.
3. The orientation histograms of 4 x 4 sample regions of each key points are used to calculate the SIFT features.

By computing the 128 dimensional SIFT descriptors at the 83 sparse feature points, we obtain a 10624 dimensional feature vector to represent each 2D facial image. To reduce the computational cost, all the SIFT feature vectors are reduced from 10624-dimensional vector space to 500-dimensional feature vector space via a common principal component analysis (PCA) transformation matrix.

3. Our Approach to Expression Recognition for Non-frontal View Facial Images

In this section, we will derive our new approach to the non-frontal view facial expression recognition. For the simplicity of the derivation, let C denote the number of the facial expression classes and let K denote the number of the views of each class.

3.1. Two-class Multi-view Facial Expression Recognition Under Bayes Theoretical Framework

Let E_{rl}^{ij} denote the Bayes error of classifying the expressions between the ith view of the rth class and the jth view of the lth class, where $r \neq l$. Then we have [4]

$$E_{rl}^{ij} = \int \min(P_{ir}p_{ir}(\mathbf{x}), P_{jl}p_{jl}(\mathbf{x}))d\mathbf{x}. \tag{1}$$

P_{ir} and P_{jl} are prior probabilities, and $p_{ir}(\mathbf{x})$ and $p_{jl}(\mathbf{x})$ are class-conditional probability density functions.

The basic idea of our method is to handle the two-class facial expression recognition problem, i.e., classifying the rth class and the lth class, as a multi-class facial expression recognition problem, where the facial images from the same view and class is viewed as an independent subclass. However, it should be noted that the Bayes error between any two subclasses that belong to the same basic expression category should be zero. The Graph illustrated in Fig. 2 shows the basic idea of our method, where the multi-class Bayes error problem was divided into several pairs of two-class Bayes error problem. Based on Figure 2, we obtain that the Bayes error between the rth class and the lth class, denoted by E_{rl}, satisfies the following inequality [2]:

$$E_{rl} \leq \sum_{i=1}^{K}\sum_{j=1}^{K} E_{rl}^{ij}. \tag{2}$$

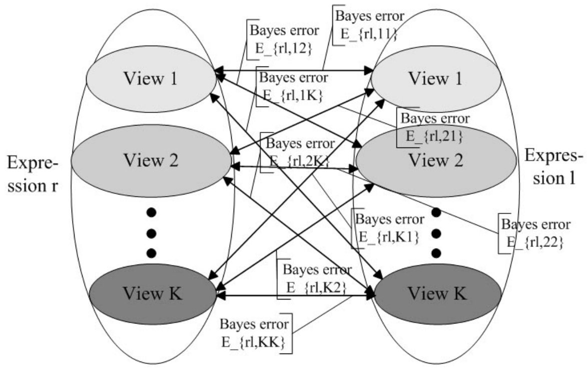

Figure 2. The Graph describing the Bayes error between the rth class and the lth class, where each class contains K different views.

If we suppose that each subclass is Gaussian distributed, i.e.,

$$p_{ir}(\mathbf{x}) = N(\mathbf{m}_{ir}, \mathbf{\Sigma}_{ir}) \tag{3}$$

then applying the following inequality

$$\min(a, b) \leq \sqrt{ab}, \quad \forall a, b \geq 0. \tag{4}$$

to the expression in (1), we obtain that the Bayes error E^{ij}_{rl} can be expressed as:

$$\begin{aligned} E^{ij}_{rl} &\leq \int \sqrt{P_{ir}P_{jl}p_{ir}(\mathbf{x})p_{jl}(\mathbf{x})}d\mathbf{x} \\ &= \sqrt{P_{ir}P_{jl}}\exp(-d^{ij}_{rl}), \end{aligned} \tag{5}$$

where d^{ij}_{rl} is the Bhattacharyya distance defined by $d^{ij}_{rl} = \frac{1}{8}(\mathbf{m}_{ir} - \mathbf{m}_{jl})^T(\bar{\mathbf{\Sigma}}^{ij}_{rl})^{-1}(\mathbf{m}_{ir} - \mathbf{m}_{jl}) + \frac{1}{2}\ln\frac{|\bar{\mathbf{\Sigma}}^{ij}_{rl}|}{\sqrt{|\mathbf{\Sigma}_{ir}||\mathbf{\Sigma}_{jl}|}}$, where $\bar{\mathbf{\Sigma}}^{ij}_{rl} = \frac{1}{2}(\mathbf{\Sigma}_{ir} + \mathbf{\Sigma}_{jl})$.

Now projecting the sample vectors to 1D feature by a vector ω, the distribution of $p_{ir}(\mathbf{x})$ becomes

$$p_{ir}(\mathbf{x}) = N(\omega^T\mathbf{m}_{ir}, \omega^T\mathbf{\Sigma}_{ir}\omega), \tag{6}$$

and the Bhattacharyya distance becomes

$$d^{ij}_{rl} = \frac{1}{8}\frac{[\omega^T(\mathbf{m}_{ir} - \mathbf{m}_{jl})]^2}{\omega^T\overline{\mathbf{\Sigma}}^{ij}_{rl}\omega} + \frac{1}{2}\ln\frac{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega}{\sqrt{(\omega^T\mathbf{\Sigma}_{ir}\omega)(\omega^T\mathbf{\Sigma}_{jl}\omega)}}. \tag{7}$$

Let $u = \omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega$, $v = \omega^T\Delta\mathbf{\Sigma}^{ij}_{rl}\omega$, where $\Delta\mathbf{\Sigma}^{ij}_{rl} = \frac{1}{2}(\mathbf{\Sigma}_{ir} - \mathbf{\Sigma}_{jl})$. Then d^{ij}_{rl} can be written as

$$\begin{aligned} d^{ij}_{rl} &= \frac{1}{8}\frac{[\omega^T(\mathbf{m}_{ir} - \mathbf{m}_{jl})]^2}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega} - \frac{1}{4}\ln(1 - (\frac{v}{u})^2) \\ &\approx \frac{1}{8}\frac{[\omega^T(\mathbf{m}_{ir} - \mathbf{m}_{jl})]^2}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega} + \frac{1}{4}(\frac{v}{u})^2, \end{aligned} \tag{8}$$

and the Bayes error E^{ij}_{rl} can be expressed as

$$\begin{aligned} E^{ij}_{rl} &\leq \sqrt{P_{ir}P_{jl}}\exp(-d^{ij}_{rl}) \\ &\approx \sqrt{P_{ir}P_{jl}}\exp\left(-\frac{1}{8}\frac{[\omega^T(\mathbf{m}_{ir} - \mathbf{m}_{jl})]^2}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega} - \frac{1}{4}(\frac{v}{u})^2\right) \\ &\approx \sqrt{P_{ir}P_{jl}} - \frac{\sqrt{P_{ir}P_{jl}}}{4}\left(\frac{\omega^T\Delta\mathbf{\Sigma}^{ij}_{rl}\omega}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega}\right)^2 \\ &- \frac{\sqrt{P_{ir}P_{jl}}}{8}\frac{[\omega^T(\mathbf{m}_{ir} - \mathbf{m}_{jl})]^2}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega}. \end{aligned} \tag{9}$$

Without loss of generality, we assume the equal prior probability for all subclasses, i.e., $P_{ir} = P$ $(i = 1, \cdots, K;$ $r = 1, \cdots, C)$, then (9) can be simplified as

$$E^{ij}_{rl} \leq P - \frac{P}{8}\frac{[\omega^T(\mathbf{m}_{ir} - \mathbf{m}_{jl})]^2}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega} - \frac{P}{4}\left(\frac{\omega^T\Delta\mathbf{\Sigma}^{ij}_{rl}\omega}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega}\right)^2. \tag{10}$$

3.2. Multi-class Multi-view Facial Expression Recognition Under Bayes Theoretical Framework

For the C-class facial expression recognition problem, we use Fig. 3 to illustrate the components of the overall Bayes error, where E_{rl} and E_{lr} denotes the same Bayes error between the rth class and the lth class. Let E denote the overall Bayes error probability of classifying the C classes expressions, then from the multi-class Bayes error theory [2] and (10) we have

$$E \leq \sum_{r=1}^{C-1}\sum_{l=r+1}^{C} E_{rl} = \frac{1}{2}\sum_{r=1}^{C}\sum_{\substack{l=1,\\ l\neq r}}^{C} E_{rl} \tag{11}$$

Combining (2), (9) and (11), we obtain that the Bayes error in (11) can be expressed as

$$\begin{aligned} E &\leq \frac{1}{2}\sum_{r}\sum_{l\neq r}\sum_{i}\sum_{j} E^{ij}_{rl} \leq \frac{1}{2}\sum_{r}\sum_{l\neq r}\sum_{i}\sum_{j} P \\ &- \frac{P}{16}\sum_{r}\sum_{l\neq r}\sum_{i}\sum_{j}\frac{[\omega^T(\mathbf{m}_{ir} - \mathbf{m}_{jl})]^2}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega} \\ &- \frac{P}{8}\sum_{r}\sum_{l\neq r}\sum_{i}\sum_{j}\left(\frac{\omega^T\Delta\mathbf{\Sigma}^{ij}_{rl}}{\omega^T\bar{\mathbf{\Sigma}}^{ij}_{rl}\omega}\right)^2 \end{aligned} \tag{12}$$

Recursively applying the following inequalities

$$\left(\frac{a}{b}\right)^2 + \left(\frac{c}{d}\right)^2 \geq \left(\frac{a+c}{b+d}\right)^2, \ \forall a, c \geq 0; b, d > 0 \tag{13}$$

$$\frac{a}{b} + \frac{c}{d} \geq \frac{a+c}{b+d}, \ \forall a, c \geq 0; b, d > 0 \tag{14}$$

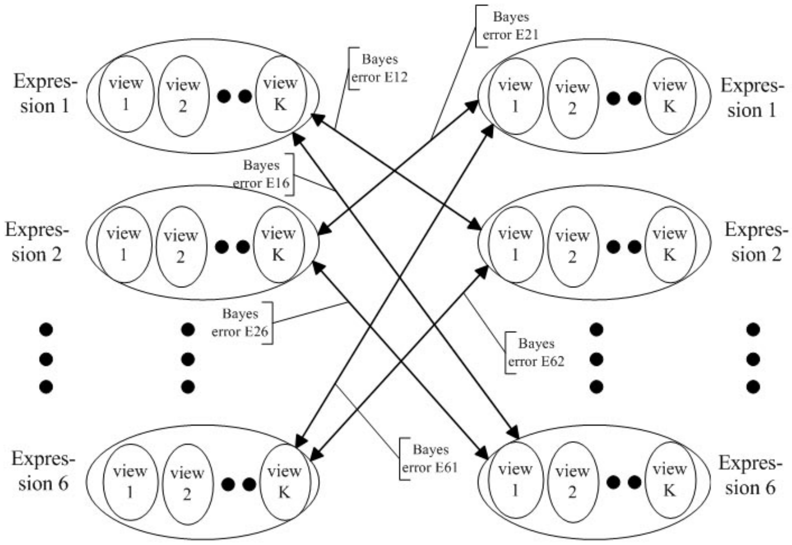

Figure 3. The Bayes error between any two different classes, where the number of the facial expressions is fixed at the 6 basic emotions. It should be noted that the Bayes error E_{rl} and E_{lr} means the same problem and thus they should be equal.

to the error bound in (12), we have

$$\begin{aligned} E &\le \frac{1}{2}\sum_{r}\sum_{l\neq r}\sum_{i}\sum_{j} P \\ &- \frac{P^3\sum_r\sum_{l\neq r}\sum_i\sum_j[\omega^T(\mathbf{m}_{ir}-\mathbf{m}_{jl})]^2}{16\sum_r\sum_{l\neq r}\sum_i\sum_j P^2\omega^T\bar{\mathbf{\Sigma}}_{rl}^{ij}\omega} \\ &- \frac{P}{8}\left(\frac{\sum_r\sum_{l\neq r}\sum_i\sum_j P^2|\omega^T\Delta\mathbf{\Sigma}_{rl}^{ij}\omega|}{\sum_r\sum_{l\neq r}\sum_i\sum_j P^2\omega^T\bar{\mathbf{\Sigma}}_{rl}^{ij}\omega}\right)^2 \quad (15) \end{aligned}$$

Let $\mathbf{B} = \sum_r\sum_{l\neq r}\sum_i\sum_j(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T$, and $\bar{\mathbf{\Sigma}} = \sum_r\sum_{l\neq r}\sum_i\sum_j P^2\bar{\mathbf{\Sigma}}_{rl}^{ij}$. Then we obtain that to minimize the Bayes error, we should minimize its upper bound, which boils down to maximizing the following discriminant criterion:

$$J(\omega) = \frac{\omega^T\mathbf{B}\omega}{\omega^T\bar{\mathbf{\Sigma}}\omega} + \frac{1}{2P^2}\left(\frac{\sum_r\sum_{l\neq r}\sum_i\sum_j P^2|\omega^T\Delta\mathbf{\Sigma}_{rl}^{ij}\omega|}{\omega^T\bar{\mathbf{\Sigma}}\omega}\right)^2 \quad (16)$$

For the ease of computation, we use the following formula as the discriminant criterion:

$$\begin{aligned} J(\omega,\mu) &= \frac{\omega^T\mathbf{B}\omega}{\omega^T\bar{\mathbf{\Sigma}}\omega} + \frac{4\mu}{2P^2}\frac{\sum_r\sum_{l\neq r}\sum_i\sum_j P^2|\omega^T\Delta\mathbf{\Sigma}_{rl}^{ij}\omega|}{\omega^T\bar{\mathbf{\Sigma}}\omega} \\ &= \frac{\omega^T\mathbf{B}\omega}{\omega^T\bar{\mathbf{\Sigma}}\omega} + \mu\frac{\sum_r\sum_{l\neq r}\sum_i\sum_j|\omega^T(\mathbf{\Sigma}_{ir}-\mathbf{\Sigma}_{jl})\omega|}{\omega^T\bar{\mathbf{\Sigma}}\omega} \end{aligned} \quad (17)$$

where $0 < 4\mu < 1 \Leftrightarrow 0 < \mu < \frac{1}{4}$ is a parameter for compensate the change of the second item in $J(\omega)$.

On the other hand, we note that

$$\begin{aligned} &\sum_{j=1}^{K}|\omega^T(\mathbf{\Sigma}_{ir}-\mathbf{\Sigma}_{jl})\omega| \ge |\sum_{j=1}^{K}\omega^T(\mathbf{\Sigma}_{ir}-\mathbf{\Sigma}_{jl})\omega| \\ &= K|\omega^T(\mathbf{\Sigma}_{ir} - \frac{1}{K}\sum_{j=1}^{K}\mathbf{\Sigma}_{jl})\omega| \quad (18) \end{aligned}$$

Let $\bar{\mathbf{\Sigma}}_l = \frac{1}{K}\sum_{j=1}^{K}\mathbf{\Sigma}_{jl}$. Then, from (17) and (18), we obtain

$$J(\omega,\mu) \ge \frac{\omega^T\mathbf{B}\omega}{\omega^T\bar{\mathbf{\Sigma}}\omega} + K\mu\frac{\sum_r\sum_{l\neq r}\sum_i|\omega^T(\mathbf{\Sigma}_{ir}-\bar{\mathbf{\Sigma}}_l)\omega|}{\omega^T\bar{\mathbf{\Sigma}}\omega}. \quad (19)$$

Let $g(\omega,\mu) = \frac{\omega^T\mathbf{B}\omega}{\omega^T\bar{\mathbf{\Sigma}}\omega} + K\mu\frac{\sum_r\sum_{l\neq r}\sum_i|\omega^T(\mathbf{\Sigma}_{ir}-\bar{\mathbf{\Sigma}}_l)\omega|}{\omega^T\bar{\mathbf{\Sigma}}\omega}$, then we obtain that maximizing criterion $g(\omega,\mu)$ will also maximize criterion $J(\omega,\mu)$. To simplify the computation, we use the criterion $g(\omega,\mu)$ as the discriminant criterion to derive the optimal discriminant vectors for the dimensionality reduction of the facial expression feature vectors.

Based on the criterion $g(\omega,\mu)$, we define the following optimal set of discriminant vectors:

$$\begin{aligned} \omega_1 &= \arg\max_{\omega} g(\omega,\mu), \\ &\cdots \\ \omega_k &= \arg\max_{\substack{\omega^T\omega_j = 0,\\ j=1,\cdots,k-1}} g(\omega,\mu) \end{aligned} \quad (20)$$

Let $\omega = \bar{\mathbf{\Sigma}}^{-\frac{1}{2}}\alpha$, $\hat{\mathbf{\Sigma}}_{ir} = \bar{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{\Sigma}_{ir}\bar{\mathbf{\Sigma}}^{-\frac{1}{2}}$, $\hat{\bar{\mathbf{\Sigma}}}_l = \bar{\mathbf{\Sigma}}^{-\frac{1}{2}}\bar{\mathbf{\Sigma}}_l\bar{\mathbf{\Sigma}}^{-\frac{1}{2}}$ and $\hat{\mathbf{B}} = \bar{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{B}\bar{\mathbf{\Sigma}}^{-\frac{1}{2}}$. Then the discriminant criterion $g(\omega, \mu)$ can be expressed as

$$\hat{g}(\alpha, \mu) \quad = \frac{\alpha^T\hat{\mathbf{B}}\alpha}{\alpha^T\alpha} + \mu\frac{\sum_r\sum_{l\neq r}\sum_i |\alpha^T(\hat{\mathbf{\Sigma}}_{ir} - \hat{\bar{\mathbf{\Sigma}}}_l)\alpha|}{\alpha^T\alpha}$$

Hence, solving the optimization problem in (20) is equivalent to solving the following optimization problem:

$$\begin{aligned} \alpha_1 &= \arg\max_{\alpha} \hat{g}(\alpha, \mu) \\ \cdots & \\ \alpha_k &= \arg \max_{\alpha^T\mathbf{U}_{k-1} = \mathbf{0}} \hat{g}(\alpha, \mu) \end{aligned} \tag{21}$$

where $\mathbf{U}_{k-1} = [\bar{\mathbf{\Sigma}}^{-1}\alpha_1, \bar{\mathbf{\Sigma}}^{-1}\alpha_2, \cdots, \bar{\mathbf{\Sigma}}^{-1}\alpha_{k-1}]$.

3.3. Solution method

Let $\mathbf{S} = [s_{rli}]$ be a $C \times C \times K$ tensor whose elements $s_{rli} \in \{+1, -1\}$. For simplicity, we call any matrix like $\mathbf{S}$ as a *sign tensor*. Now we define a matrix $\mathbf{T}(\mathbf{S}, \mu)$ associated with $\mathbf{S}$ and μ as:

$$\mathbf{T}(\mathbf{S}, \mu) \quad = \quad \hat{\mathbf{B}} + \mu\sum_r\sum_{l\neq r}\sum_i s_{rli}(\hat{\mathbf{\Sigma}}_{ir} - \hat{\bar{\mathbf{\Sigma}}}_l)$$

Let $\mathbf{\Omega} = \{\mathbf{S}|\mathbf{S} = [s_{rli}], s_{rli} \in \{+1, -1\}\}$ denote the *sign tensor* set. Then we obtain that

$$\hat{g}(\alpha, \mu) = \max_{\mathbf{S}\in\mathbf{\Omega}}\{\frac{\alpha^T\mathbf{T}(\mathbf{S}, \mu)\alpha}{\alpha^T\alpha}\} \tag{22}$$

From (21) and (22), we obtain that the optimal vectors α_i $(i = 1, 2, \cdots, k)$ can be expressed as

$$\begin{aligned} \alpha_1 &= \arg\max_{\mathbf{S}\in\mathbf{\Omega}}\max_{\alpha}\frac{\alpha^T\mathbf{T}(\mathbf{S}, \mu)\alpha}{\alpha^T\alpha} \\ \cdots & \\ \alpha_k &= \arg\max_{\mathbf{S}\in\mathbf{\Omega}}\max_{\alpha^T\mathbf{U}_{k-1}=\mathbf{0}}\frac{\alpha^T\mathbf{T}(\mathbf{S}, \mu)\alpha}{\alpha^T\alpha} \end{aligned} \tag{23}$$

3.3.1 Solving the first discriminant vector α_1

To solve the first vector α_1 in (23), it is crucial to find the optimal sign tensor $\mathbf{S}$ associated with α_1. If the sign tensor $\mathbf{S}$ is fixed, then the optimal discriminant vector is the eigenvector associated with the largest eigenvalue of the matrix $\mathbf{T}(\mathbf{S}, \mu)$. Solving the principal eigenvector of $\mathbf{T}(\mathbf{S}, \mu)$ can be efficiently realized via the power iteration approach [5].

Necessary Condition 1: Suppose that $\mathbf{S}$ is the optimal sign tensor and α is the associated principal eigenvector of $\mathbf{T}(\mathbf{S}, \mu)$. Let s_{rli} be the element of $\mathbf{S}$ in the $r \times l \times i$ position. Then the following holds because the sign of both s_{rli} and $\alpha^T(\hat{\mathbf{\Sigma}}_{ir} - \hat{\bar{\mathbf{\Sigma}}}_l)\alpha$ are the same:

$$s_{rli}\alpha^T(\hat{\mathbf{\Sigma}}_{ir} - \hat{\bar{\mathbf{\Sigma}}}_l)\alpha \geq 0.$$

Definition 1: Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be two sign tensors and α_1 and α_2 are the associated principal eigenvector of $\mathbf{T}(\mathbf{S}_1, \mu)$ and $\mathbf{T}(\mathbf{S}_2, \mu)$, respectively. If $\alpha_1^T\mathbf{T}(\mathbf{S}_1, \mu)\alpha_1 > \alpha_2^T\mathbf{T}(\mathbf{S}_2, \mu)\alpha_2$, then we say that $\mathbf{S}_1$ is better than $\mathbf{S}_2$.

Suppose that $\mathbf{S}_1^*$ is the sign tensor associated with the optimal vector α_1. Then from definition 1 and (23), we obtain that $\mathbf{S}_1^*$ is better than any sign tensor $\mathbf{S} \in \mathbf{\Omega}$. Consequently, finding the optimal sign tensor $\mathbf{S}_1^*$ is the process of finding the best sign tensor in $\mathbf{\Omega}$.

In what follows, we propose a greedy approach to find the suboptimal sign tensor $\mathbf{S}$ and then find a suboptimal vector α_1. The basic idea of this approach is to fix a proper value ($+1$ or -1) for each element of $\mathbf{S}$. We show our approach as the following steps:

1. Set an initial value for $\mathbf{S}$, e.g, $s_{rli} \leftarrow -1$;
2. Solve the principal eigenvector of $\mathbf{T}(\mathbf{s}, \mu)\alpha = \lambda\alpha$ via power iteration method, and set $\lambda_0 \leftarrow \lambda$;
3. For $r = 1, 2, \cdots, C$, $l = 1, 2, \cdots, C$, and $i = 1, 2, \cdots, K$ Do
    - Set $s_{rli} \leftarrow -s_{rli}$;
    - Solve the principal eigenvector of $\mathbf{T}(\mathbf{S}, \mu)\alpha = \lambda\alpha$ via power iteration method, and set $\lambda_1 \leftarrow \lambda$;
    - If $\lambda_1 \leq \lambda_0$, then $s_{rli} \leftarrow -s_{rli}$, else $\lambda_0 \leftarrow \lambda_1$;

After performing the above steps, we can obtain a better sign tensor $\mathbf{S}$. Moreover, we can repeat steps 2 to 3 until the sign parameter vector $\mathbf{S}$ converges.

3.3.2 Solving the $(k + 1)$-th discriminant vector α_{k+1}

Suppose that we have obtained the first k vectors $\alpha_1, \cdots, \alpha_k$. To solve the $(k + 1)$-th vector α_{k+1}, we introduce the following lemma and theorems [12].

Lemma 1. Let $\mathbf{Q}$ be a $d \times p$ $(p < d)$ matrix with orthonormal columns. If $\alpha^T\mathbf{Q} = \mathbf{0}$, then there exists a (non-unique) $\beta \in \mathbb{R}^d$ such that $\alpha = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\beta$.

Proof: We can find the complement basis $\mathbf{Q}^\perp$ such that the matrix $\tilde{\mathbf{Q}} = (\mathbf{Q} \quad \mathbf{Q}^\perp)$ is an orthogonal matrix. Then we have $\mathbf{Q}^\perp(\mathbf{Q}^\perp)^T = \mathbf{I}_d - \mathbf{Q}\mathbf{Q}^T$ due to $\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^T = \mathbf{I}_d$. From $\alpha^T\mathbf{Q} = \mathbf{0}$, there exists a $\gamma \in \mathbb{R}^{d-p}$ such that $\alpha = \mathbf{Q}^\perp\gamma$. On the other hand, $\text{rank}\{(\mathbf{Q}^\perp)^T\} = d - p$. Therefore, the columns of $(\mathbf{Q}^\perp)^T$ form a basis of $\mathbb{R}^{d-p}$. So there exists a $\beta \in \mathbb{R}^d$, such that $\gamma = (\mathbf{Q}^\perp)^T\beta$. Thus, we have $\alpha = \mathbf{Q}^\perp(\mathbf{Q}^\perp)^T\beta = (\mathbf{I}_d - \mathbf{Q}\mathbf{Q}^T)\beta$.

Theorem 1. Let $\mathbf{Q}_r\mathbf{R}_r$ be the QR decomposition of $\mathbf{U}_r$, where $\mathbf{R}$ is an $r \times r$ upper triangular matrix. Then α_{r+1} defined in (23) is the principal eigenvector corresponding to the largest eigenvalue of the following matrix $(\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{T}(\mathbf{S}, \mu)(\mathbf{I}_d - \mathbf{Q}_k\mathbf{Q}_r^T)$.

Proof: Since $\alpha^T\mathbf{U}_r = \mathbf{0}$, $\mathbf{Q}_r\mathbf{R}_r$ is the QR decomposition of $\mathbf{U}_r$, and $\mathbf{R}_r$ is non-singular, we obtain that

$\alpha^T\mathbf{Q}_r = \mathbf{0}$. From Lemma 1, there exists a $\beta \in \mathbb{R}^d$ such that $\alpha = (\mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^T)\beta$. Moreover, it should be noted that $(\mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^T)(\mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^T) = \mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^T$. Thus, we have $\alpha = (\mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^T)\beta = (\mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^T)\alpha$. Hence, we obtain that $\max_{\alpha^T\mathbf{U}_r=\mathbf{0}} \frac{\alpha^T\mathbf{T}(\mathbf{S},\mu)\alpha}{\alpha^T\alpha} = \max_\alpha \frac{\alpha^T(\mathbf{I}-\mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{T}(\mathbf{S},\mu)(\mathbf{I}-\mathbf{Q}_r\mathbf{Q}_r^T)\alpha}{\alpha^T\alpha}$. Therefore, α_{r+1} is the principal eigenvector corresponding to the largest eigenvalue of the matrix $(\mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{T}(\mathbf{S},\mu)(\mathbf{I} - \mathbf{Q}_r\mathbf{Q}_r^T)$.

Theorem 2. Suppose that $\mathbf{Q}_r\mathbf{R}_r$ is the QR decomposition of $\mathbf{U}_r$. Let $\mathbf{U}_{r+1} = (\mathbf{U}_r \quad \alpha_{r+1})$, $\mathbf{q} = \alpha_{r+1} - \mathbf{Q}_r(\mathbf{Q}_r^T\alpha_{r+1})$, and $\mathbf{Q}_{r+1} = \begin{pmatrix} \mathbf{Q}_r & \frac{\mathbf{q}}{\|\mathbf{q}\|} \end{pmatrix}$. Then $\mathbf{Q}_{r+1}\begin{pmatrix} \mathbf{R}_r & \mathbf{Q}_r^T\alpha_{r+1} \\ \mathbf{0} & \|\mathbf{q}\| \end{pmatrix}$ is the QR decomposition of $\mathbf{U}_{r+1}$.

Proof: From $\mathbf{q} = \alpha_{r+1} - \mathbf{Q}_r(\mathbf{Q}_r^T\alpha_{r+1})$ and the fact that $\mathbf{Q}_r\mathbf{R}_r$ be the QR decomposition of $\mathbf{U}_r$, we obtain that $\mathbf{Q}_r^T\mathbf{q} = \mathbf{0}$ and $\mathbf{Q}_r^T\mathbf{Q}_r = \mathbf{I}_r$, where $\mathbf{I}_r$ is the $r \times r$ identity matrix. Thus, we have

$$\mathbf{Q}_{r+1}^T\mathbf{Q}_{r+1} = \begin{pmatrix} \mathbf{Q}_r^T\mathbf{Q}_r & \mathbf{Q}_r^T\frac{\mathbf{q}}{\|\mathbf{q}\|} \\ \frac{\mathbf{q}^T\mathbf{Q}_r}{\|\mathbf{q}\|} & 1 \end{pmatrix} = \mathbf{I}_{r+1} \qquad (24)$$

where $\mathbf{I}_{r+1}$ is the $(r+1) \times (r+1)$ identity matrix. On the other hand,

$$\begin{pmatrix} \mathbf{Q}_r & \frac{\mathbf{q}}{\|\mathbf{q}\|} \end{pmatrix}\begin{pmatrix} \mathbf{R}_r & \mathbf{Q}_r^T\alpha_{r+1} \\ \mathbf{0} & \|\mathbf{q}\| \end{pmatrix} = (\mathbf{Q}_r\mathbf{R}_r \quad \alpha_{r+1})$$
$$= (\mathbf{U}_r \quad \alpha_{r+1}) = \mathbf{U}_{r+1}. \qquad (25)$$

From (24) and (25), one can see that the theorem is true.

Based on the above two theorems, we can solve (23) in an efficient way: Theorem 1 makes it possible to use the power method to solve (23) and Theorem 2 makes it possible to update $\mathbf{Q}_{r+1}$ from $\mathbf{Q}_r$ by adding a single column. Moreover, we have

$$\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T = (\mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)(\mathbf{I}_d - \mathbf{q}_r\mathbf{q}_r^T) \qquad (26)$$

where $\mathbf{q}_r$ is the rth column of $\mathbf{Q}_r$. Equation (26) makes it possible to use the ROU technique for fast updating the positive semidefinite matrix $(\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{T}(\mathbf{S}_{r+1},\mu)(\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T)$ from $(\mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)\mathbf{T}(\mathbf{S}_{r+1},\mu)(\mathbf{I}_d - \mathbf{Q}_{r-1}\mathbf{Q}_{r-1}^T)$. We give the pseudo-code of solving these k discriminant vectors of the proposed method in Algorithm 1.

4. Experiments

In order to validate our algorithm, we conduct facial expression recognition experiments with the multi-view face images of 100 subjects that we have generated from the BU-3DFE database, as described in Section 2. We compare our algorithm with PCA [4] and Fisher's LDA [1] based on the

Algorithm 1: Solving the optimal vectors ω_i ($i = 1, 2, \cdots, k$)

```
Input:
  • Data matrix $\mathbf{X}$, class label vector $\mathbf{L}$, and parameter $\mu$.

Initialization:
  1. Compute the covariance matrices $\boldsymbol{\Sigma}_{ir}$ and $\bar{\boldsymbol{\Sigma}}_r$ ($r = 1, \cdots, C$ and $i = 1, \cdots, K$), $\bar{\boldsymbol{\Sigma}}$, and $\mathbf{B}$;
  2. Perform SVD of $\bar{\boldsymbol{\Sigma}}$: $\bar{\boldsymbol{\Sigma}} = \mathbf{U}\Lambda\mathbf{U}^T$, compute $\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}} = \mathbf{U}\Lambda^{-\frac{1}{2}}\mathbf{U}^T$ and $\bar{\boldsymbol{\Sigma}}^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}^T$;
  3. Compute $\hat{\boldsymbol{\Sigma}}_{ir} = \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{ir}\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ ($r = 1, \cdots, C$; $i = 1, \cdots, K$) and $\hat{\mathbf{B}} = \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{B}\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$, and let $C = \max(\mathbf{L})$;

For $i = 1, 2, \cdots, k$, Do
  1. Set $\mathbf{S} \leftarrow$ ones$(C, C, K)$, $\mathbf{S}_1 \leftarrow -\mathbf{S}$, and $\mathbf{T}(\mathbf{S},\mu) \leftarrow \hat{\mathbf{B}} + \mu\sum_r\sum_{l\neq r}\sum_i s_{rli}(\hat{\boldsymbol{\Sigma}}_{ir} - \hat{\bar{\boldsymbol{\Sigma}}}_l)$;
  2. Solve the principal eigenvector of $\mathbf{T}(\mathbf{S},\mu)\alpha = \lambda\alpha$ via power iteration method, and set $\lambda_0 \leftarrow \lambda$;
     While $\mathbf{S} \neq \mathbf{S}_1$, Do
       (a) Set $\mathbf{S}_1 \leftarrow \mathbf{S}$;
       (b) For $r, l = 1, 2, \cdots, C$; $i = 1, \cdots, K$, Do
             • Set $s_{rli} \leftarrow -s_{rli}$ and $\mathbf{T}(\mathbf{S},\mu) \leftarrow \hat{\mathbf{B}} + \mu\sum_r\sum_{l\neq r}\sum_i s_{rli}(\hat{\boldsymbol{\Sigma}}_{ir} - \hat{\bar{\boldsymbol{\Sigma}}}_l)$;
             • Solve the principal eigenvector of $\mathbf{T}(\mathbf{S},\mu)\alpha = \lambda\alpha$ via power iteration method, and set $\lambda_1 \leftarrow \lambda$;
             • If $\lambda_1 \leq \lambda_0$, then $s_{rli} \leftarrow -s_{rli}$, else $\lambda_0 \leftarrow \lambda_1$;
       (c) Compute $\mathbf{T}(\mathbf{S},\mu) = \hat{\mathbf{B}} + \mu\sum_r\sum_{l\neq r}\sum_i s_{rli}(\hat{\boldsymbol{\Sigma}}_{ir} - \hat{\bar{\boldsymbol{\Sigma}}}_l)$ and solve the principal eigenvector of $\mathbf{T}(\mathbf{S},\mu)\alpha_i = \lambda\alpha_i$ via power iteration method;
  3. If $i = 1$, $\mathbf{q}_i \leftarrow \alpha_i$, $\mathbf{q}_i \leftarrow \mathbf{q}_i/\|\mathbf{q}_i\|$, and $\mathbf{Q}_1 \leftarrow \mathbf{q}_i$; else $\mathbf{q}_i \leftarrow \alpha_i - \mathbf{Q}_{i-1}(\mathbf{Q}_{i-1}^T\alpha_i)$, $\mathbf{q}_i \leftarrow \mathbf{q}_i/\|\mathbf{q}_i\|$, and $\mathbf{Q}_i \leftarrow (\mathbf{Q}_{i-1} \quad \mathbf{q}_i)$;
  4. Compute $\hat{\bar{\boldsymbol{\Sigma}}}_p \leftarrow \hat{\bar{\boldsymbol{\Sigma}}}_p - (\hat{\bar{\boldsymbol{\Sigma}}}_p\mathbf{q}_i)\mathbf{q}_i^T - \mathbf{q}_i(\mathbf{q}_i^T\hat{\bar{\boldsymbol{\Sigma}}}_p) + \mathbf{q}_i(\mathbf{q}_i^T\hat{\bar{\boldsymbol{\Sigma}}}_p\mathbf{q}_i)\mathbf{q}_i^T$ ($p = 1, \cdots, C$);
  5. Compute $\hat{\bar{\boldsymbol{\Sigma}}}_{pq} \leftarrow \hat{\bar{\boldsymbol{\Sigma}}}_{pq} - (\hat{\bar{\boldsymbol{\Sigma}}}_{pq}\mathbf{q}_i)\mathbf{q}_i^T - \mathbf{q}_i(\mathbf{q}_i^T\hat{\bar{\boldsymbol{\Sigma}}}_{pq}) + \mathbf{q}_i(\mathbf{q}_i^T\hat{\bar{\boldsymbol{\Sigma}}}_{pq}\mathbf{q}_i)\mathbf{q}_i^T$ ($p = 1, \cdots, C$; $q = 1, \cdots, K$ );
  6. Compute $\hat{\mathbf{B}} \leftarrow \hat{\mathbf{B}} - \hat{\mathbf{B}}\mathbf{q}_i\mathbf{q}_i^T - \mathbf{q}_i(\mathbf{q}_i^T\hat{\mathbf{B}}) + \mathbf{q}_i(\mathbf{q}_i^T\hat{\mathbf{B}}\mathbf{q}_i)\mathbf{q}_i^T$

Output:
  • $\omega_i = \frac{\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}}{\sqrt{\alpha_i^T\bar{\boldsymbol{\Sigma}}^{-1}\alpha_i}}\alpha_i$, $i = 1, 2, \cdots, k$.
```

same K-Nearest Neighbor (KNN) classifier. More specifically, our experiments are carried out as follows: For each experiment, we run 10 independent trials. In each trial, we randomly partition the 100 subjects into two groups. One group contains face images of 80 subjects of all 6 expressions, 4 intensities, and 5 views and comprises a training set of 9600 face images. The other group contains faces images of 20 subjects of all 6 expressions, 4 intensities, and 5 views and comprises a test set of 2400 face images. Each trial of an experiment involves a different random partition of the 100 subjects into a training set and a test set, and the results of the 10 independent trials are averaged.

Table 1 shows the overall error rates of the three methods and Fig. 4 shows the overall confusion matrix of recognizing the six expressions using our method, where the overall error rate of each recognition method is obtained by averaging all the error rates tested on all the views and expressions of testing images. From Table 1, we can see that the lowest overall error rate is achieved as low as 21.65% by using our method, which is much better than the result obtained by Hu et al. [6] (=33.5%). From Fig. 4 we can see that, among the six expressions, happy expression and surprise expression are easier to be recognized whereas fear expression and angry expression are more difficult to be recognized.

Table 1. The overall error rates (%) of various methods.

	Our method	Fisher's LDA	PCA
Error rate (%)	21.65	23.10	33.84

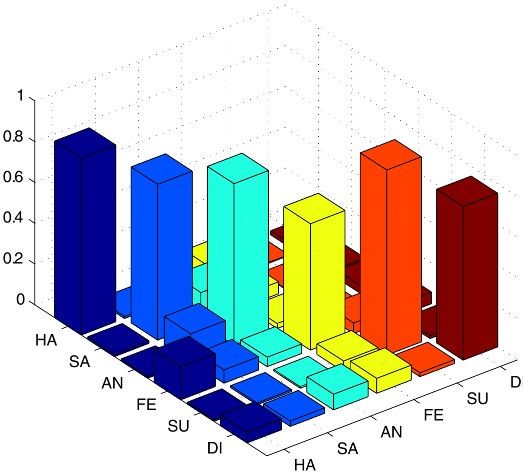

Figure 4. The overall confusion matrix of our method.

To compare the recognition performance with respect to different facial views, we plot the average error rate versus the different facial views using the three methods in Fig. 5. From Fig. 5 we can see that, for each facial view, the average error rate of our method is lower than the other two methods. We can also see from Fig. 5 that, among the various views, the best recognition performance is achieved when the facial views are between 30^o and 60^o. Moreover,

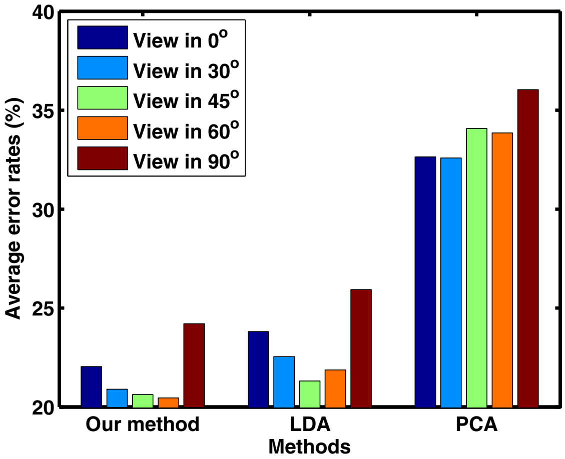

Figure 5. The average error rates of six expressions with the different choices of views using the different recognition methods.

we also show the average error rate (%) of different emotions versus different views using our method in Table 2, from which we can see the different recognition results of each expression with the change of the views and the best results are achieved when the facial views are between 30^o and 60^o.

Table 2. Average error rate (%) of different emotions versus different views using our method.

	0^o	30^o	45^o	60^o	90^o	Ave
Hap	12.50	12.50	12.87	**11.88**	16.87	13.32
Sad	27.87	23.00	**22.00**	21.13	24.25	23.65
Ang	19.75	19.12	**18.75**	19.63	27.12	20.88
Fear	38.25	38.38	38.12	**35.75**	39.75	38.05
Sur	8.88	9.12	9.00	**8.87**	10.50	**9.27**
Dist	25.00	23.25	**23.00**	25.50	26.75	24.70

Moreover, to compare the recognition performance of the three methods in recognizing each expression across all the views, we plot the average error rate of each expression across all the views in Fig. 6. From Fig. 6 we can see again that, for all the three recognition methods, the happy expression and the surprise expression are the easier expressions to be recognized whereas the fear expression is the most difficult expression to be recognized.

Finally, we investigate the influences of the dimensionality of the reduced features on the recognition performance

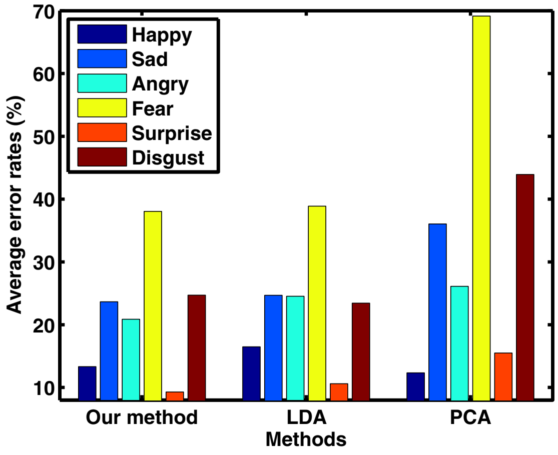

Figure 6. The average error rates of each expression across all the views using the different recognition methods.

of our method. For this purpose, we show the overall recognition results versus the different choices of the number of reduced features under the various views in Fig. 7. It can be clearly seen from Fig. 7 that the error rate of recognizing the expressions in each facial view decrease with the increase of the number of projection features until the number of projection features reach 25 or so. After that, the error rates become insensible to the increase of the projection features.

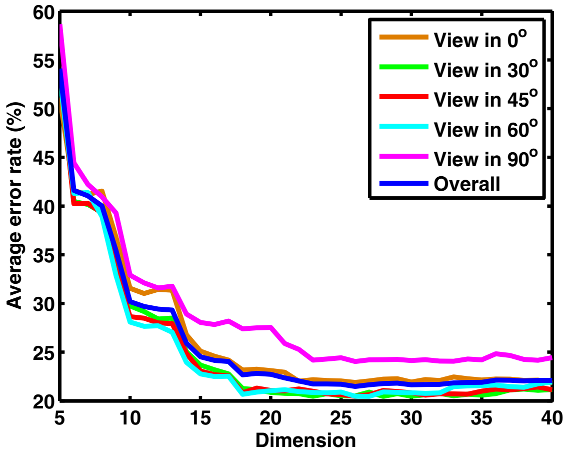

Figure 7. Average error rates versus the number of features selected by our method.

5. Conclusions

In this paper we developed a novel theory of multi-class classification, based on minimizing an estimated closed-form Bayes error for an important, difficult, and much less studied problem, i.e., the non-frontal facial expression recognition. The extensive experimental on the BU-3DFE database showed that the non-frontal view face images can achieve better recognition rates than the frontal view face images, especially when the facial views fell into the region between 30^o and 60^o, where the lowest error rate (= 21.65%) is obtained.

The major limitation of the proposed method is that the 83 ground truth points we used as the key points for SIFT feature extraction was provided by the BU-3DFE database rather than automatically located by computer. For our future work, we consider to use the real-time facial point detecting and tracking techniques to solve this problem.

Acknowledgment

This work was partly supported by National Natural Science Foundation of China under Grants 60503023 and 60872160, and partly supported by the US Government VACE program.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, 19(7):711–720, 1997. 6

[2] J. T. Chu and J. C. Chuen. Error probability in decision functions for character recognition. *Journal of the Association for Computing Machinery*, 14(2):273–280, 1967. 2, 3

[3] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003. 1

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition (Second Edition)*. Academic Press, New York. 2, 6

[5] G. Golub and C. Van. *Matrix Computations*. The Johns Hopkins University Press, 1996. 1, 5

[6] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T. Huang. A study of non-frontal-view facial expressions recognition. In *Proceedings of ICPR*, pages 1–4, 2008. 1, 7

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[8] Y. L. Tian, T. Kanade, and J. F. Cohn. Facial expression analysis. In *S. Z. Li, A. K. Jain (Eds.), Handbook of Facial Recognition*, pages 247–276, New York, USA, 2005. Springer. 1

[9] J. Wang, L. Yin, X. Wei, and Y. Sun. 3d facial expression recognition based on primitive surface feature distribution. In *Proceedings of CVPR*, pages 1399–1406, 2006. 1

[10] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3d facial expression database for facial behavior research. In *Proceedings of 7th International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006. 1, 2

[11] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE TPAMI*, 31(1):39–58, 2009. 1

[12] W. Zheng. Heteroscedastic feature extraction for texture classification. *IEEE Signal Processing Letters*, 16(9):766–769, 2009. 5