




# Sparse modified marginal fisher analysis for facial expression recognition

Zhe Wang<sup>1</sup> · Li Zhang<sup>1,2</sup>  · Bangjun Wang<sup>1,2</sup>

Published online: 30 January 2019  
© Springer Science+Business Media, LLC, part of Springer Nature 2019

## Abstract

Marginal Fisher analysis (MFA) is an efficient method for dimension reduction, which can extract useful discriminant features for image recognition. Since sparse learning can achieve better generalization ability and lessen the amount of computations in recognition tasks, this paper introduces sparsity into MFA and proposes a novel sparse modified MFA (SMMFA) method for facial expression recognition. The goal of SMMFA is to extract discriminative features by using the resulted sparse projection matrix. First, a modified MFA is proposed to find the original projection matrix. Similar to MFA, the modified MFA also defines the intra-class graph and the inter-class graph to describe geometry structure in the same class and local discriminant structure between different classes, respectively. In addition, the modified MFA removes the null space of the total scatter matrix. The sparse solution of SMMFA can be gained by solving the  $\ell_1$ -minimization problem on the original projection matrix using the linearized Bregman iteration. Experimental results show that the proposed SMMFA can effectively extract intrinsic features and has better discriminant power than the state-of-the-art methods.

**Keywords** Facial expression recognition · Marginal fisher analysis · Sparse learning · Linearized Bregman iteration

## 1 Introduction

Over the past few decades, on account of its potential applications in various fields such as psychology, lie-detection, human-computer interaction (HCI) and psychopathology analysis, facial expression recognition (FER) has been a quite active and crucial research topic in the realm of affective computing [1–5]. The primary job of FER is using a set of labelled training facial expression images to label the unlabelled test images. Extracting expression features using effective and efficient facial image description approaches is the major task of constructing automatic facial expression recognition systems. In terms of feature extraction, approaches for FER can be broadly classified into two chief categories [1]. One of them is geometric-based methods while the other is appearance-based methods. Geometric-based methods accomplish a given

recognition task by describing a facial geometric structure, that is, facial component shapes such as the edge of eyes and mouth, based on a series of salient landmarks. Active appearance model (AAM) [6] and Ekman's facial action coding system (FACS) [7, 8] are the two most typical geometric-based approaches. FACS is based on forty four action units (AUs) [9], which can be roughly interpreted as the smallest visible units of human's facial muscular motions. One problem of geometric-based methods which cannot be ignored is that the recognition performance of those methods mainly depends on the accurate detection of facial components which is fairly tough and not reliable in automatic FER systems [1]. The recent research on FER has demonstrated that the appearance-based approaches are more powerful than geometric-based approaches in term of achieving higher recognition accuracy [5, 10–12].

Among appearance-based approaches for FER, most methods focus on subspace learning [13–15]. As the facial images are always represented as high-dimensional vectors and some of the measured variables in high-dimensional data are redundant which are not useful for understanding the underlying relevant information, dimension reduction becomes a vital and essential task in FER [5]. Facilitating the visualization and interpretation of original data, dimension reduction algorithms project the original high-dimensional data into a

---

✉ Li Zhang  
zhangliml@mail.suda.edu.cn

<sup>1</sup> School of Computer Science and Technology & Joint International Research Laboratory of Machine Learning and Neuromorphic Computing, Soochow University, Suzhou 215006, Jiangsu, China

<sup>2</sup> Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, Jiangsu, China

low-dimensional subspace in which the intrinsic and useful information in expression images would be retained. Common used methods include principal component analysis (PCA) [16–18], locality preserving projection (LPP) [19, 20], and linear discriminant analysis (LDA) [21–23].

PCA and LDA are the two typical dimension reduction approaches which have been widely utilized in FER. PCA is an unsupervised method which accomplishes the task of dimension reduction by finding a few orthogonal linear combinations of the original variables with the largest variances. Compared with PCA, LDA is a supervised method and tries to maximize the ratio of between-class scatter to within-class scatter. The core idea of LDA is quite simple, that is, samples in the same class should cluster tightly together, while samples from different classes are always as far as possible from each other in the lower-dimensional representation. The application of LDA in FER is limited for the following two reasons. One is that the dimension of the sample space is quite larger than the number of samples in the training set in theory, that is, the so-called small sample size (SSS) problem [24], which enhances the difficulty of implementing robust and fast recognition. The other is that LDA characterizes the discriminant capability of between-class and within-class, which is optimal only in the cases that data for each class approximately obeys the Gaussian distribution [25]. To overcome the limitations of LDA, lots of variants have been proposed, such as null space LDA [26], LDA/GSVD [27] and LDA/QR [28]. Yan et al. proposed the Marginal Fisher Analysis (MFA) [25], which was developed by using the graph embedding framework as a platform to extract features. In MFA, two graphs are designed to characterize the within-class compactness and the between-class separability, respectively. The number of available projection directions for MFA is fairly greater than that of LDA, overcoming the limitation of LDA that the number of available projection directions is fewer than the number of the classes. In addition, there is no prior assumption on the data distribution in MFA, which makes it more reliable in real world applications than LDA.

Sparse learning has been proved to be a prominent modeling tool to implement the task of dimension reduction effectively by obtaining models of high-dimensional data with high degree of interpretability. Moreover, sparse projection matrices could lessen the amount of computations and be stored efficiently, which makes sparse learning approaches more robust and practical in constructing automatic recognition system. The goal of sparsity is to minimize the empirical loss using as few as features as possible. The most natural way of imposing sparsity is to penalize the objective with the  $\ell_0$ -norm, which is mostly NP-hard and difficult to be solved directly. The  $\ell_1$ -norm regularization, which is effective in avoiding overfitting, is also an alternative way to approximate the  $\ell_0$ -norm and obtain sparse solution [29]. The least absolute shrinkage and selection operator (LASSO) [30] is such a

method using the  $\ell_1$  approximation of the  $\ell_0$ -norm penalty. Similar to LASSO, basis pursuit uses the  $\ell_1$ -norm to replace the  $\ell_0$ -norm in order to solve optimization problems [31], which can also be used to obtain sparse solutions.

The projection matrices of the dimension reduction algorithms mentioned above are not sparse, which means each feature in the low-dimensional space is a linear combination of all the features of the original data and the coefficients of the linear combination are generally non-zero. This makes the interpretation of the extracted features difficult when the data dimension is large. To overcome this drawback, sparse principal component analysis (SPCA) using LASSO to produce sparse principal components [32] and sparse linear discriminant analysis (SLDA) learning sparse discriminant subspace [33] are proposed. The sparse solutions of SLDA and SPCA are both found by utilizing sparse regression. If the regularization parameters are zeros, SLDA and SPCA can derive the exact solutions of the original problem, but the obtained solutions are not sparse. And if the parameter of the  $\ell_1$ -norm penalty is larger than zero, these sparse solutions are the approximate solutions of the original objective function. Both of SPCA and SLDA enhance the generalization ability and provide the psychological and physiological interpretation. Wang et al. proposed a sparse local Fisher discriminant analysis (SLFDA) method to recognize facial expression, obtaining the sparse solution by solving the basis pursuit problem derived from solutions of the local Fisher discriminant analysis (LFDA). The projection matrix of SLFDA is sparse, which can make the physical meaning of the extracted features clear and procure preferable effects [5]. Puthenputhussery et al. proposed a sparse representation model using the so-called complete marginal fisher analysis framework [12], that is CMFA-SR for short, and applied the proposed method to visual recognition tasks. However, CMFA-SR utilizes sparse representation to derive the final sparse features for recognition tasks instead of a sparse projection matrix.

To generate a sparse projection matrix, this paper also introduces sparsity into MFA and proposes a novel sparse subspace learning method, or sparse modified MFA (SMMFA). First, a modified MFA (MMFA) is proposed to find the original projection matrix. Similar to MFA, MMFA constructs the intra-class adjacency graph and inter-class adjacency graph to describe the local geometric information of samples in the same class and the local discriminant information of samples from different classes, respectively. Moreover, MMFA removes the null space of the total scatter matrix. Inspired by SLFDA, we obtain the sparse solution to SMMFA by finding the minimum  $\ell_1$ -norm solution from the solution of the modified MFA. When the minimum  $\ell_1$ -norm solution can be formulated as an  $\ell_1$ -minimization problem, we can use the linearized Bregman iteration to handle the  $\ell_1$ -minimization problem. The major contributions of this paper contain the following points:

- (1) First, this paper proposes a modified MFA to generate a non-sparse original projection matrix for extracting more effective discriminant features. Compared to MFA, MMFA removes the information in the null space of the total scatter matrix, which is not related with the discriminant ability.
- (2) Second, this paper presents a sparse MMFA (SMMFA) to generate a sparse projection matrix through applying the linear Bregman iteration to the  $\ell_1$ -minimization problem which is related to the original projection matrix generated by MMFA. The sparsity of the projection matrix makes SMMFA achieve better generalization ability. It is the first time for MFA-based method to consider getting a sparse projection matrix.

The remainder of this paper is organized as follows. Section 2 describes the proposed SMMFA method in detail. Section 3 discusses the connections of SMMFA with other related work. Section 4 shows and analyses experimental results of the presented method. Section 5 concludes the paper.

## 2 Sparse modified MFA

In this section, we address the proposed sparse modified MFA in detail. The framework of SMMFA is given in Fig. 1, where  $\mathbf{P}$  is the original projection matrix generated by the modified MFA which removes the null space in the total scatter matrix  $\mathbf{S}$ ,  $\mathbf{V}$  is the sparse projection matrix obtained by applying the linearized Bregman iteration to  $\mathbf{P}$ , and  $\otimes$  denotes the operation of matrix multiplication. The training dataset is used to learn the original projection matrix  $\mathbf{P}$ . The framework of SMMFA can be simply described as follows.

First, SMMFA has two adjacency graphs, the intra-class graph and the inter-class graph to describe the local geometry structure in the same class and the local discriminant structure between different classes, respectively, which is similar to MFA. Different from MFA, SMMFA defines a total scatter matrix and removes the null space of it to extract more discriminant features, which will be discussed in detail in the following section 2.1. Second, the linearized Bregman iteration method is used to obtain the sparse solution of SMMFA

by solving the  $\ell_1$ -minimization problem on the solution gained before. Finally, the labels of test data could be predicted by the classification model trained in the projected subspace.

### 2.1 Modified MFA

This section presents the modified MFA. Assume that we have a set of  $N$  samples:  $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ , where  $\mathbf{x}_i \in \mathbb{R}^D$ ,  $y_i \in \{1, 2, \dots, c\}$  is the label of  $\mathbf{x}_i$ ,  $N$  is the number of samples,  $D$  is the dimension of each sample, and  $c$  is the number of classes. MMFA defines two adjacency graphs, the intra-class graph and the inter-class graph. The elements of the intra-class graph  $\mathbf{F}^w \in \mathbb{R}^{N \times N}$  are defined as follows:

$$F_{ij}^w = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \pi_K^+(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \pi_K^+(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where  $\pi_K^+(\mathbf{x}_j)$  denotes the set of  $K$  homogenous nearest neighbours of  $\mathbf{x}_j$ ,  $\mathbf{x}_i \in \pi_K^+(\mathbf{x}_j)$  means that  $\mathbf{x}_i$  is the homogenous nearest neighbour of  $\mathbf{x}_j$ , and  $\mathbf{x}_i$  and  $\mathbf{x}_j$  belong to the same class. The elements of the inter-class graph  $\mathbf{F}^b \in \mathbb{R}^{N \times N}$  are defined as follows:

$$F_{ij}^b = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \pi_K^-(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \pi_K^-(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where  $\pi_K^-(\mathbf{x}_j)$  denotes the set of  $K$  heterogeneous nearest neighbors of  $\mathbf{x}_j$ , and  $\mathbf{x}_i$  and  $\mathbf{x}_j$  have different class labels.

Similar to MFA, MMFA is to maximize the inter-class scatter to extract marginal discriminant information between different classes, and minimize the intra-class scatter to extract the local similarity information in the same class. In this way, MMFA could retain the geometry structure of data.

The intra-class scatter is

$$\sum_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2 F_{ij}^w$$

and the inter-class scatter is

$$\sum_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2 F_{ij}^b$$

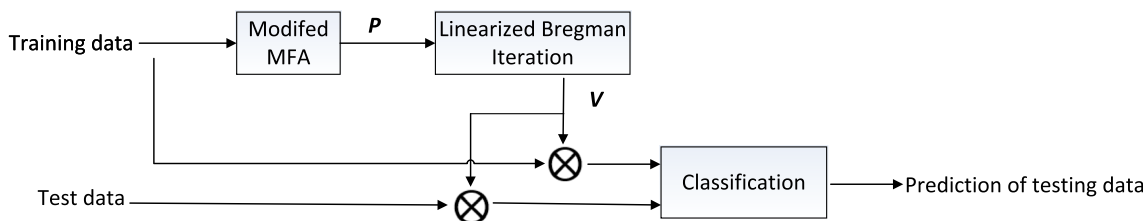


Fig. 1 Framework of SMMFA, where  $\mathbf{P}$  is the projection matrix generated by using the modified MFA,  $\mathbf{V}$  is the sparse projection matrix, and  $\otimes$  denotes matrix multiplication

Obviously, the intra-class scatter matrix has the form:

$$S_w = XL_wX^T \tag{3}$$

and the inter-class scatter matrix is defined as:

$$S_b = XL_bX^T \tag{4}$$

where  $X \in \mathbb{R}^{D \times N}$  is the sample matrix composed of training data, the Laplacian matrices  $L_w = D^w - F^w \in \mathbb{R}^{N \times N}$  and  $L_b = D^b - F^b$ , both  $D^w$  and  $D^b$  are diagonal matrices, and  $D_{ii}^w = \sum_j F_{ij}^w$  and  $D_{ii}^b = \sum_j F_{ij}^b$ .

Then, the total scatter matrix is  $S_t = X(L_w + L_b)X^T$ . It was found that the null space in the total scatter matrix  $S_t$  cannot be utilized to enhance the discriminant power and can be removed [5, 34]. Thus, we adopt this viewpoint and introduce it to MFA, which leads to the modified MFA.

Suppose that the total scatter matrix  $S_t$  can be decomposed by

$$S_t = A\Lambda A^T \tag{5}$$

where  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$  is the eigenvalue matrix with  $\lambda_i > 0, i = 1, 2, \dots, r$ , the rank  $r = \text{rank}(S_t)$ ,  $A = [a_1, a_2, \dots, a_r] \in \mathbb{R}^{D \times r}$  is the matrix consisting of eigenvectors of  $S_t$  corresponding to  $r$  positive eigenvalues. For both  $S_w$  and  $S_b$ , we need to remove the null space of  $S_t$  from them.

Let  $m$  be the size of the subspace or the projected subspace, and  $P \in \mathbb{R}^{D \times m}$  be the projection matrix of the modified MFA. Then, the objection function of MMFA is given as followed:

$$\max_P \frac{P^T (A^T S_b A) P}{P^T (A^T S_w A) P} \tag{6}$$

Assume that  $\overline{F}_b = A^T S_b A$  and  $\overline{F}_w = A^T S_w A$ , then the solution  $P$  to (6) can be gained by performing eigendecomposition on  $(\overline{F}_w)^{-1} \overline{F}_b$ . Let  $p_i$  be eigenvectors.

Then the projection matrix can be represented  $P = [p_1, p_2, \dots, p_m]$  consisting of  $m$  eigenvectors corresponding to the first  $m$  largest eigenvalues.

### 2.2 Reducing computation

The dimension  $D$  of the samples is always quite high, which makes it difficult to directly calculate the eigenvectors of the  $D \times D$  matrix  $S_t$ . For computational considerations, we need to improve the efficiency of calculating  $A$ .

First, we can get the intra-class graph  $F^w$  and the inter-class graph  $F^b$  using (1) and (2), respectively. Then, we have the global graph  $F^t = F^b + F^w$ , and the total Laplacian matrix  $L_t = L_w + L_b$ . According to the spectral graph theory in [35], the total Laplacian matrix  $L_t$  is symmetric and positive semi-definite. Thus, we can decompose  $L_t$  as

$$L_t = Q_t \Lambda_t Q_t^T \tag{7}$$

where  $Q_t$  is the orthogonal eigenvector matrix and  $\Lambda_t = \text{diag}(\lambda_t^1, \lambda_t^2, \dots, \lambda_t^N)$ ,  $\lambda_t^i \geq 0$  is the eigenvalue matrix of  $L_t$ .

We can construct an auxiliary matrix and perform matrix decomposition on it instead of  $S_t$  to reduce computational complexity. Assume that  $S_t = XL_tX^T = XQ_t\Lambda_tQ_t^T X^T = H_t H_t^T$ , the auxiliary matrix  $H_t$  can be constructed by

$$H_t = XQ_t\Lambda_t^{\frac{1}{2}} \tag{8}$$

Then, we can perform the thin singular value decomposition on  $H_t \in \mathbb{R}^{D \times N}$ , and have

$$H_t = A\overline{\Lambda}_t Q \tag{9}$$

where  $A$  is the left singular vector matrix,  $\overline{\Lambda}_t$  is the singular value matrix and  $Q$  is the right singular vector matrix of  $H_t$ . Note that the left singular vector matrix  $A = [a_1, a_2, \dots, a_r] \in \mathbb{R}^{D \times r}$  is the one that we need, which is the eigenvector matrix of  $S_t$ .

By performing the thin singular value decomposition on  $H_t$ , we can get a more efficient way to obtain the eigenvector matrix  $A$  since the size of  $H_t$  is apparently smaller than the that of  $S_t$ .

### 2.3 Sparse projection matrix

Generally, the solution to (6) is not sparse. A sparse projection matrix can reduce the computational complexity of algorithms and be stored efficiently. Thus, SMMFA aims at getting the sparse projection matrix based on the solution of MMFA. The sparse solution is gained by the sparsification of the original projection matrix  $P$ .

Let  $V \in \mathbb{R}^{D \times m}$  be the sparse projection matrix. Then the sparse projection matrix can be found by solving the following  $\ell_1$ -norm minimization problems:

$$\min_V \|V\|_1 \tag{10}$$

*s.t.*  $A^T V = P$

where  $\|V\|_1 = \sum_{i=1}^D \sum_{j=1}^m |v_{ij}|$ , and  $A$  is the eigenvectors of  $S_t$ .

Actually, (10) is a basis pursuit problem. Here, we use the linearized Bregman iteration [31, 36–38] to solve the basis pursuit problem, which has been considered as one of the most successful methods for solving (10). Therefore, we have the following the iteration:

$$\begin{cases} B^{k+1} = B^k - A(A^T V^k - P) \\ V^{k+1} = \xi \Gamma_\mu(B^{k+1}) \end{cases} \tag{11}$$

where  $\mathbf{B}^0 = \mathbf{V}^0 = \mathbf{0}$ ,  $\xi > 0$  is a user-defined parameter,  $\Gamma_\mu(\mathbf{B}) = [\Gamma_\mu(B_{ij})]_{ij}$  and

$$\Gamma_\mu(B_{ij}) = \begin{cases} 0, & \text{if } |B_{ij}| \leq \mu \\ \text{sgn}(B_{ij}) (|B_{ij}| - \mu), & \text{if } |B_{ij}| \geq \mu \end{cases} \quad (12)$$

where  $\text{sgn}$  is the sign function and  $\mu > 0$  is a user-defined parameter.

By iteratively calculating (11), we can obtain the sparse projection matrix  $\mathbf{V}$ . Algorithm 1 shows the detail procedure of SMMFA.

When the sparse projection matrix  $\mathbf{V}$  is obtained, we can project the training samples into the subspace spanned by  $\mathbf{V}$ . Let  $\mathbf{X}' = \mathbf{V}^T \mathbf{X}$  be the projected matrix of training data. Then, we can adopt the nearest neighbour classifier with the Euclidean distance for the classification tasks.

### 3 Connection to other related work

In this section, we briefly describe some related work, including MFA and CMFA-SR.

#### 3.1 MFA

Since the inter-class scatter in LDA cannot effectively characterize the separability of different classes in the case of data without the Gaussian distribution, MFA was proposed to characterize the intra-class compactness and the inter-class separability without the Gaussian distribution assumption [25]. MFA defines the intra-class scatter matrix and inter-class scatter matrix, as (3) and (4) shown in Section 2.1. The goal of MFA is to make samples in the same class closer with each other and samples in different classes farther as possible in the projected space. Thus, the objective function of MFA can be defined as

$$\max_P \frac{\mathbf{P}^T \mathbf{S}_b \mathbf{P}}{\mathbf{P}^T \mathbf{S}_w \mathbf{P}} \quad (13)$$

where  $\mathbf{P}$  is the projection matrix.

Note that projection matrix  $\mathbf{P}$  gained by MFA is not sparse, which is different from SMMFA. What's more, compared with MFA, SMMFA defines a total scatter matrix  $\mathbf{S}_t$  and removes its null space to enhance the discriminant power.

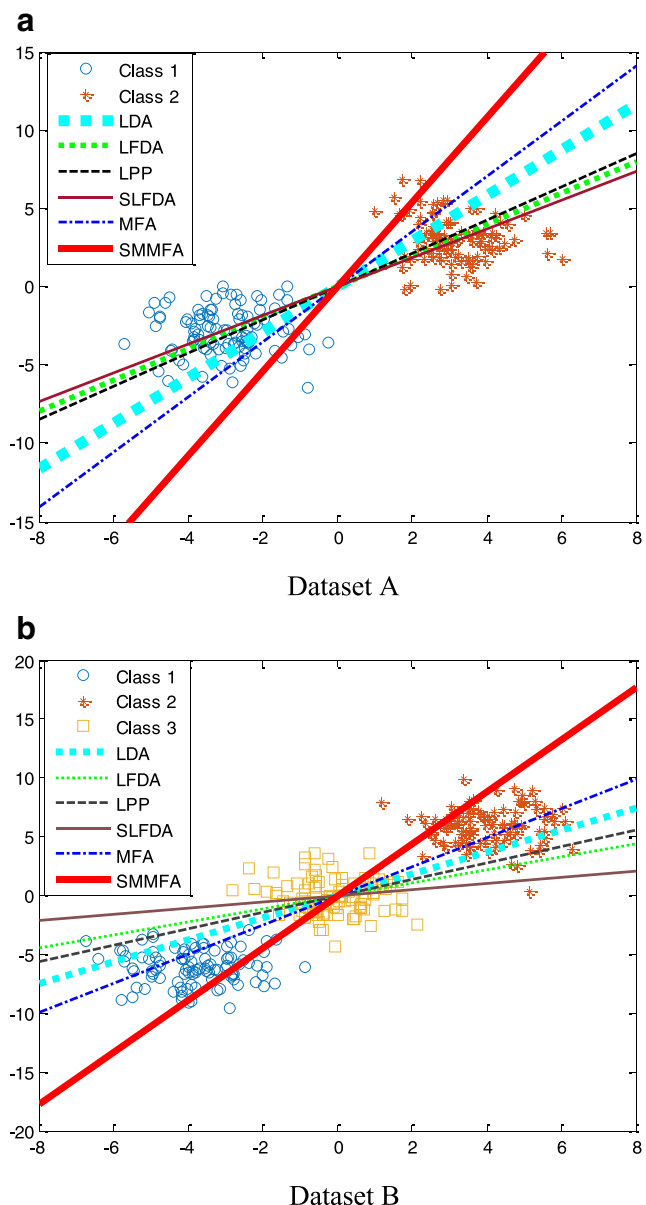


Fig. 2 Comparison of projection directions obtained by six methods on toy Dataset A (a) and Dataset B (b)

Table 1 Average ratio of between-class scatter to the within-class scatter in the projected subspace

	Dataset A	Dataset B
None	2.7902 ± 0.0503	3.2414 ± 0.1176
LDA	4.6757 ± 0.2302	5.0307 ± 0.2016
LFDA	4.3866 ± 0.1632	4.7884 ± 0.1811
LPP	4.4747 ± 0.1477	4.7758 ± 0.2146
SLFDA	4.3930 ± 0.2198	4.0905 ± 0.2595
MFA	4.7191 ± 0.2635	5.0462 ± 0.2530
SMMFA	4.7476 ± 0.2545	5.1348 ± 0.2415

**Algorithm 1.** Sparse Modified Marginal Fisher Analysis

**Input:** The training sample matrix  $X = [x_1, x_2, \dots, x_N]$ , the corresponding label vector  $Y = [y_1, y_2, \dots, y_N]$ ,  $y_i$  is the label of  $x_i$ ,  $N$  is the number of samples, the parameters  $\xi$  and  $\mu$ , and the tolerance  $\varepsilon$ .

**Output:** The sparse projection matrix  $V$ .

1. Construct the intra-class graph  $F^w$  using (1) and calculate the intra-class scatter matrix  $S_w$  using (3);
2. Construct the inter-class graph  $F^b$  using (2) and calculate the intra-class scatter matrix  $S_b$  using (4);
3. Construct the total Laplacian matrix  $L_t$  and perform eigen-decomposition on it to obtain  $Q_t$  and  $\Lambda_t$ .
4. Generate  $H_t$  using (8) and perform the thin singular value decomposition on  $H_t$  as (9) to obtain  $A$ .
5. Solve the generalized eigenvalue problem (6) to obtain the original projection matrix  $P$ ;
6. Let  $B^0 = V^0 = 0$  and  $k = 1$ .
7. Calculate  $V^{k+1}$  using (11) repeatedly and let  $k = k + 1$  until  $\|A^T V^{k+1} - P\|_F < \varepsilon$ .
8. Return  $V = V^{k+1}$ .

**3.2 CMFA-SR**

CMFA-SR is an algorithm based on MFA and uses sparse representation to gain the sparse features for classification tasks [12]. CMFA-SR first utilizes the complete MFA algorithm to extract features in both the column space of the local samples-based within-class scatter matrix and the null space of its transformed matrix. After that, CMFA-SR makes the extracted features have sparsity.

The main differences between SMMFA and CMFA-SR lie in three aspects. First, the core idea of SMMFA is to learn a sparse projection matrix to derive discriminant features, while the goal of CMFA-SR is to obtain sparse discriminant features instead of a sparse projection matrix. Second, SMMFA uses a modified MFA to generate the original projection matrix  $P$ , and CMFA-SR adopts the complete MFA to implement the projection task, which extracts features in two subspaces. Thus, the projection procedures of these two methods are different. Third, the projected features of SMMFA can be directly used for classification tasks, while the projected features gained by CMFA-SR still need to be transformed to sparse

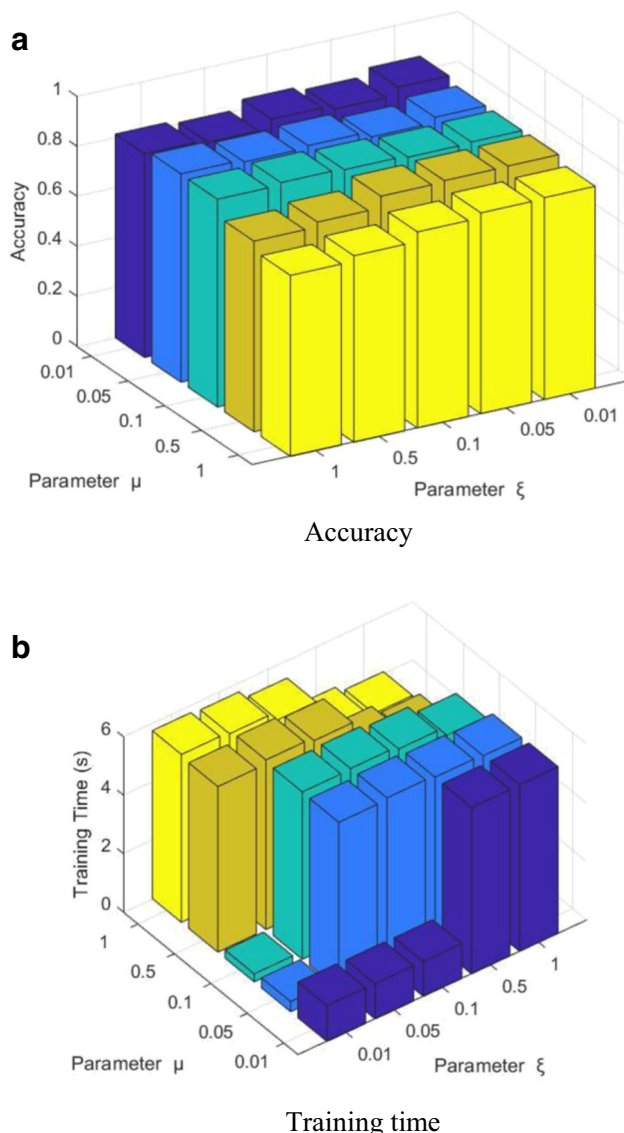


**Fig. 3** Cropped facial expression images from the JAFFE facial expression database with respect to seven expressions: ‘happy’ (the first column), ‘sadness’ (the second column), ‘surprise’ (the third column), ‘anger’ (the fourth column), ‘disgust’ (the fifth column), ‘fear’ (the sixth column) and ‘neutral’ (the seventh column)

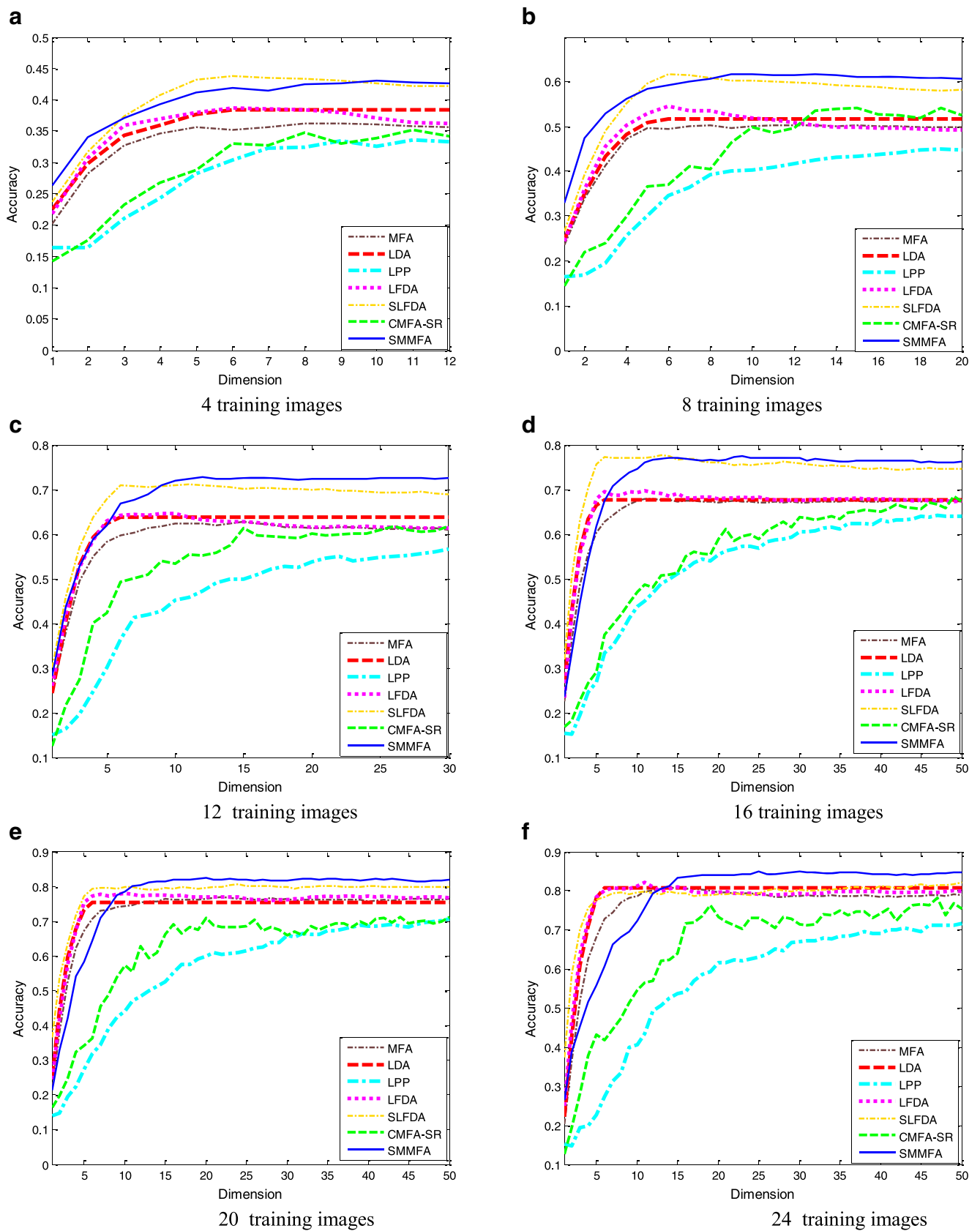
features. The procedure of SMMFA is simpler than that of CMFA-SR. Obviously, the computational complexity of CMFA-SR is much greater than that of SMMFA since each training sample or test sample need to be projected and sparsification in CMFA-SR.

**4 Experiments**

To demonstrate the effectiveness of SMMFA, we perform experiments on two toy data sets and three public FER databases: the Japanese Female Facial Expression (JAFFE) database [39], the Extended Cohn-Kanade dataset (CK+) [40] and the GEMEP-FERA database [41]. The nearest neighbour classifier with the Euclidean distance is adopted to implement all



**Fig. 4** Accuracy (a) and training time (b) vs. parameters  $\xi$  and  $\mu$  in SMMFA



**Fig. 5** Average accuracy vs. dimension on the JAFFE database under different training images in each facial expression class. (a) 4 training images. (b) 8 training images. (c) 12 training images. (d) 16 training images. (e) 20 training images. (f) 24 training images

**Table 2** Comparison of best average accuracy and standard deviation (%) on the JAFFE database

	4 train	8 train	12 train	16 train	20 train	24 train
LPP	33.54 ± 2.66(11)	44.94 ± 4.75(19)	56.63 ± 3.73(30)	64.16 ± 4.80(47)	70.82 ± 6.44(50)	71.67 ± 6.99(50)
LDA	38.35 ± 4.50(6)	51.72 ± 4.19(6)	63.76 ± 5.30(6)	67.72 ± 6.71(6)	75.41 ± 4.88(6)	80.78 ± 4.80(6)
MFA	36.27 ± 4.61(9)	50.32 ± 4.83(11)	62.67 ± 3.92(15)	67.87 ± 4.93(11)	76.92 ± 6.23(22)	80.89 ± 6.30(15)
LFDA	38.73 ± 5.13(6)	54.59 ± 4.04(6)	64.65 ± 4.24(10)	69.70 ± 5.55(11)	78.01 ± 4.88(10)	82.01 ± 6.18(11)
SLFDA	<b>43.84</b> ± 4.67(6)	61.72 ± 3.94(7)	71.05 ± 4.70(11)	<b>77.57</b> ± 3.60(13)	80.62 ± 5.87(24)	82.11 ± 6.22(50)
CMFA-SR	35.23 ± 4.11(11)	54.28 ± 5.72(19)	61.78 ± 4.85(30)	68.43 ± 5.54(49)	71.11 ± 4.31(11)	78.17 ± 5.27(47)
SMMFA	43.03 ± 4.86(10)	<b>61.82</b> ± 5.62(10)	<b>72.71</b> ± 3.95(12)	77.38 ± 4.64(23)	<b>82.47</b> ± 5.54(20)	<b>84.89</b> ± 4.92(30)

The bolded values in Table 2 are the highest ones among compared methods

classification tasks. LPP, LDA, MFA, LFDA are the most widely used dimension reduction methods which have been proved to be effective and efficient in many circumstances [19, 25]. Similar to SMMFA, SLFDA generates the sparse solution by finding the minimum  $\ell_1$ -norm solution from the solution of relatively traditional dimension reduction method [5]. CMFA-SR is another algorithm based on MFA and utilizes a different sparse solution from SMMFA, achieving remarkable recognition rate in multiple vision recognition tasks [12]. All of the methods mentioned above (LPP, LDA, MFA, LFDA, SLFDA and CMFA-SR) are implemented and compared with SMMFA.

All numerical experiments are performed on a personal computer with a 3.6GHz Intel(R) Core(TM) i7-7700 and 8G bytes of memory. This computer runs Windows 7, with Matlab R2012a and VC++ 14.0 installed.

#### 4.1 Experiments on toy data sets

In order to evaluate the projection effects of the subspace learned by SMMFA, we generate two toy data sets, which are two-class (Dataset A) and three-class (Dataset B) two-dimensional Gaussian distributions, respectively. Each class has 100 data points which are randomly generated according to its distribution.

We compare just six methods, including LPP, LDA, MFA, LFDA, SLFDA and SMMFA and show their projection directions. Since CMFA-SR has more than one projection matrix and aims at sparse discriminant features, we do not consider it here. For six methods, we project the two-dimensional data points in Datasets A and B into a one-dimensional subspace, respectively.

In the Dataset A, we generate two separated Gaussian distributions, representing two-class data points. Figure 2a shows the projection vectors of six methods. We can see that all six methods can project the two-class data points into the one-dimensional subspace and obtain a good separability since the original data is linearly separable.

In the Dataset B, we generate three separated Gaussian distributions to evaluate the projected effects of the methods further. Figure 2b gives the effectiveness of six methods. Although the projected data is not linearly separable, these points are piecewise linear separable. Obviously, SMMFA has the best separability, followed by MFA and LDA.

In order to show the effectiveness of methods more explicitly, we repeat 10 experiments on two toy data sets and calculate the average ratio of the between-class scatter to the within-class scatter in the projected subspace. Note that the greater the ratio is, the better the separability is. The average ratios and their standard deviations are shown in Table 1, where “None” means the situation without using dimension reduction methods. The ratios on the toy Dataset A support the results in Fig. 2a. All six methods have good separability. Even so, SMMFA has the best separability among them, followed by MFA. The ratios on the toy Dataset B also support the results in Fig. 2b, or SMMFA has the best separability among six methods.

In a nutshell, SMFA has the best discriminant power among the six methods on the two toy sets.

#### 4.2 Experiments on JAFFE database

The JAFFE database consists of 213 facial expression images of 10 Japanese women, where each person has one neutral

**Table 3** Comparison of training time (s) on the JAFFE database

SMMFA	MFA	LDA	LPP	LFDA	SLFDA	CMFA-SR
0.3669 ± 0.0552	0.3574 ± 0.0373	0.9095 ± 0.0514	0.0764 ± 0.0129	0.4495 ± 0.0478	14.8467 ± 0.4994	3.1275 ± 0.2051



(NE) expression and six basic facial expressions, i.e., anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU). In all the 213 images, there are 29 to 32 images categorized to every basic facial expression or neutral expression. The initial image size in the JAFFE database is  $256 \times 256$ . In experiments, we manually align, crop and resize them to  $32 \times 32$  images. Figure 3 shows some aligned, cropped and resized images from the JAFFE database.

The cropped images are randomly divided into two subsets, training set and test set. The images in the training set are utilized to learn the projection subspace spanned by the eigenvectors generated by dimension reduction algorithms. Any image in the test set can be projected into the embedding subspace by using obtained projection matrices.

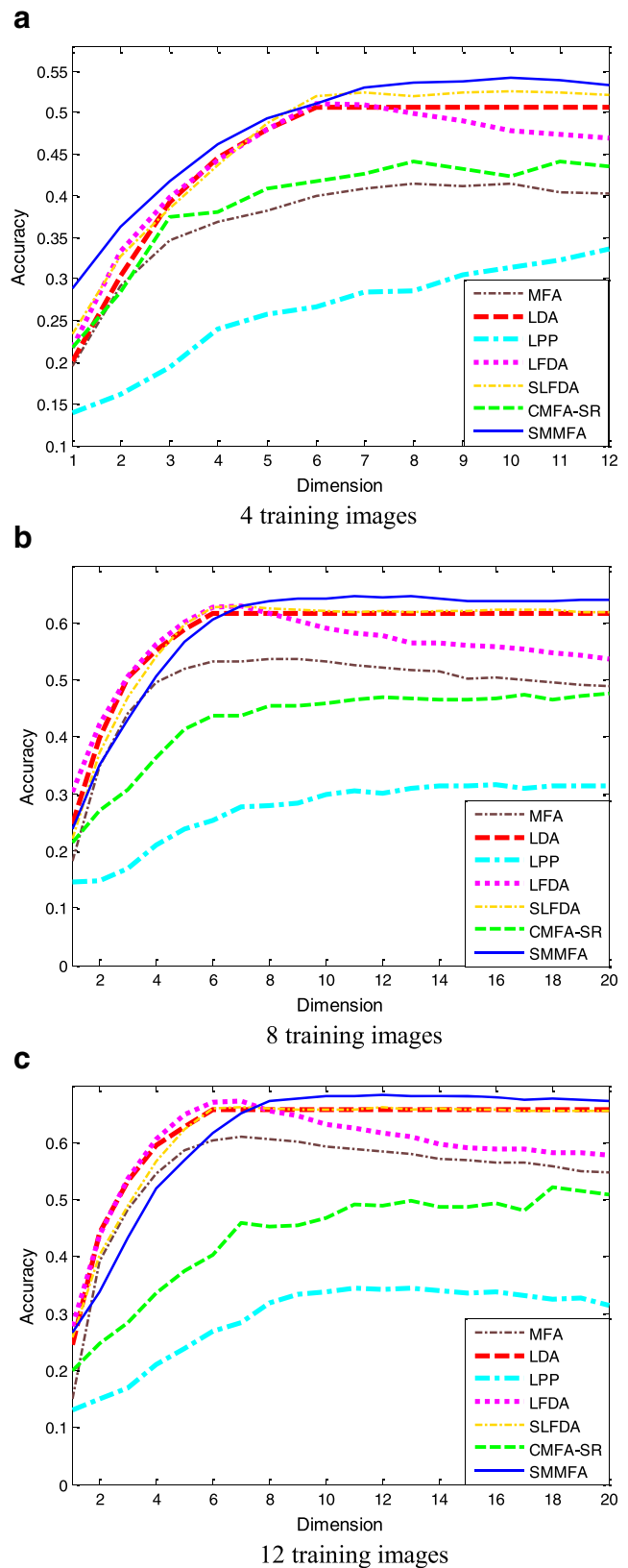
### 4.2.1 Parameter sensitivity analysis

To look into the influence of parameters  $\xi$  and  $\mu$  of SMMFA, we randomly select 20 images from each expression category of the JAFFE database and repeat the experiment for twenty times using a different setting of these two parameters. Figure 4 shows the average accuracy and average training time of SMMFA when  $\xi$  and  $\mu$  range between 0.01 and 1, respectively. From Fig. 4a, we can draw a conclusion that the recognition accuracy of SMMFA is slightly better when the value of parameter  $\xi$  or  $\mu$  is smaller. From Fig. 4b, we know that the training time of SMMFA significantly increases when the product of  $\xi$  and  $\mu$  is greater than 0.001. Moreover, the training time does not change significantly and remains a higher value when the product of  $\xi$  and  $\mu$  is between 0.001 and 0.25. When the product of  $\xi$  and  $\mu$  is greater than 0.25, the training time would slightly decrease. The shortest training time is obtained when  $\xi = 0.01$  and  $\mu$  is 0.05 or 0.1.

Based on the conclusions above, the following parameter settings are used: the values of parameters  $\xi$  and  $\mu$  in SMMFA are set to 0.01 and 0.05, respectively.



**Fig. 6** Facial expression images constructed from the CK+ database with respect to seven expressions: ‘anger’ (the first column), ‘contempt’ (the second column), ‘disgust’ (the third column), ‘fear’ (the fourth column), ‘happiness’ (the fifth column), ‘sadness’ (the sixth column) and ‘surprise’ (the seventh column)



**Fig. 7** Average recognition rate (%) vs. the dimension on the CK+ database under different training images in each facial expression class. (a) 4 training images. (b) 8 training images. (c) 12 training images

#### 4.2.2 Performance comparison

Here, we compare SMMFA with the other six methods (LPP, LDA, MFA, LFDA, SLFDA and CMFA-SR) when changing the dimension of the subspace and the number of training samples. We randomly select  $s$  images from each expression class in the JAFFE database for training, the remaining images are used for test, where  $s \in \{4, 8, 12, 16, 20, 24\}$ . In order to reduce the accidental error, for each value of  $s$ , we repeat 20 times with different training and test sets.

The variation of accuracy vs. dimensions is shown in Fig. 5. For all seven methods, the accuracy increases remarkably at first when the feature dimension increases from a certain small value, that is 1, to be precise. After that, when the feature dimension continues to increase, the recognition rates of LPP and CMFA-SR keep increasing slowly, and the recognition rates of LFDA, SLFDA, LDA, MFA and SMMFA decrease slightly after reaching their maxima. It is worth noting that when feature dimension is at a relatively small value, the recognition rate of SMMFA is often lower than some of the other approaches like SLFDA and LDA. However, SMMFA achieves a better performance than others when more features are considered.

From Fig. 5, we can also draw a conclusion that the recognition rate increases when the number of training samples increases for all compared methods here. When the number of training samples is small, see Fig. 5a–d, SMMFA and SLFDA have compared performance and are always better than other five methods. In Fig. 5e, f, SMMFA is much better than the other six methods.

The best average accuracy of each approach with the standard deviation and the corresponding optimal dimension is given in Table 2, where the bolded values are the highest ones among compared methods. As we can see from Table 2, the sparse learning methods such as SLFDA and SMMFA are more powerful than non-sparse methods such as MFA and LFDA, which shows that the sparse learning algorithms can learn the intrinsic and useful information in facial images well. Moreover, SMMFA gets the best performance as it reaches the highest recognition rate among all the implemented approaches.

#### 4.2.3 Computational complexity

For SMMFA, the computational complexity of constructing intra-class graph  $F^w$  and inter-class graph  $F^b$  in Algorithm 1 is  $O(N^2)$ , while the time complexity of obtaining matrix  $A$  is  $O(N^3)$ . Assume that the iterative times is  $t$ . Then, the computational complexity of the linear Bregman iteration process is  $O(tN^3)$ . Therefore, the overall time complexity of SMMFA is  $O(tN^3)$ .

Now, we compare the computational complexity of seven methods. Simply, we take the running time of training to

**Table 4** Comparison of best average accuracy and standard deviation (%) on the CK+ database

	4 train	8 train	12 train
LPP	33.67 ± 7.43(12)	31.75 ± 5.71(16)	34.50 ± 5.25(11)
LDA	50.65 ± 3.71(6)	61.74 ± 4.37(6)	65.79 ± 2.99(6)
MFA	41.54 ± 4.46(8)	53.70 ± 5.39(8)	61.03 ± 5.04(7)
LFDA	51.13 ± 4.45(6)	62.96 ± 4.27(7)	67.06 ± 4.39(6)
SLFDA	52.54 ± 8.71(10)	62.93 ± 8.19(7)	66.30 ± 9.64(7)
CMFA-SR	44.21 ± 5.55(8)	47.62 ± 5.83(20)	52.14 ± 6.17(18)
SMMFA	<b>54.19</b> ± 6.74(10)	<b>64.63</b> ± 6.56(11)	<b>68.44</b> ± 5.23(12)

The bolded values in Table 4 are the highest ones among compared methods

approximate the computational complexity of seven methods. We repeat 20 experiments, in which there are 20 images randomly selected from each expression category for training.

Table 3 shows the average running time of the training process for all methods when the best average accuracies in Table 2 are obtained. From Table 3, we can see that LFDA and MFA take similar training time, less than LPP and LDA does. According to Table 3, SMMFA takes slightly more time to complete the training process compared with MFA. Although SLFDA has a similar solution for finding the sparse projection matrix to SMMFA, SLFDA takes evidently more time to converge compared with LFDA, which means the convergence of SMMFA is efficient. As another method based on MFA, CMFA-SR utilizes a different sparse solution from SLFDA and SMMFA. It converges faster than SLFDA but slower than SMMFA.

In a nutshell, SMMFA still has an advantage on computational complexity among the seven approaches.



**Fig. 8** Facial expression sample images constructed from the GEMEP-FERA facial expression database with respect to five expressions: ‘anger’ (the first row), ‘fear’ (the second row), ‘joy’ (the third row), ‘relief’ (the fourth row), and ‘sadness’ (the fifth row)

**Table 5** Comparison of best average accuracy and standard deviation (%) on the GEMEP-FERA database

	4 train	8 train	12 train	16 train	20 train	24 train
LPP	35.04 ± 3.97(12)	38.06 ± 3.45(10)	43.35 ± 3.69(30)	44.14 ± 4.62(25)	46.28 ± 4.31(25)	49.64 ± 5.11(29)
LDA	39.98 ± 5.16(6)	46.42 ± 4.01(20)	51.81 ± 3.70(29)	55.18 ± 2.94(30)	57.50 ± 2.68(30)	59.94 ± 3.62(30)
MFA	37.36 ± 5.23(6)	46.10 ± 4.94(7)	52.10 ± 3.70(10)	56.18 ± 2.78(11)	58.25 ± 2.86(11)	61.11 ± 3.39(14)
LFDA	40.68 ± 4.61(5)	46.27 ± 3.63(5)	51.50 ± 3.07(5)	55.50 ± 2.74(15)	58.48 ± 2.92(17)	61.53 ± 3.73(18)
SLFDA	41.32 ± 5.79(4)	43.98 ± 7.04(15)	48.15 ± 5.84(8)	53.32 ± 4.96(29)	53.95 ± 3.88(27)	55.92 ± 5.48(30)
CMFA-SR	37.42 ± 4.22(15)	40.86 ± 3.78(28)	44.52 ± 4.25(30)	50.30 ± 4.42(28)	55.95 ± 3.10(45)	59.42 ± 3.99(50)
SMMFA	<b>41.48 ± 2.89(8)</b>	<b>47.63 ± 3.15(14)</b>	<b>53.06 ± 3.16(22)</b>	<b>56.36 ± 2.74(18)</b>	<b>59.53 ± 2.87(12)</b>	<b>62.03 ± 3.37(30)</b>

The bolded values in Table 5 are the highest ones among compared methods

### 4.3 Experiments on CK+ database

The extended Cohn-Kanade (CK+) contains 593 facial expression sequences from 123 subjects. Each sequence is categorized to one of the seven basic expressions: anger, contempt, disgust, fear, happy, sadness and surprise. The initial size of images in every sequence is 640 × 490 or 640 × 480. In this experiment, we select 314 images by cutting out the last one or two images of some sequences in CK+ database. There are 44 anger, 18 contempt, 57 disgust, 17 fear, 69 happiness, 27 sadness, and 82 surprise images in the selected image subset. All 314 images are manually cropped and resized to the resolution of 32 × 32. Some images are shown in Fig. 6. We can see that the facial images in the extended CK+ database are quite different in the lighting condition.

Similar to the experiments in the JAFFE database, we randomly select  $s$  images from each expression class for training and the remaining images are used for test, where  $s \in \{4, 8, 12\}$ . For each value of  $s$ , the experiment with different training and test is repeated 20 times as well.

The variation of average accuracy vs. dimensions with different training sample number is shown in Fig. 7. Obviously, SMMFA has the best classification performance among seven methods under three situations. The best average recognition rate of each approach, its standard deviation and the corresponding optimal dimension are given in Table 4, where the bolded values are the best ones among compared methods. The difference in the lighting condition can lead to the overall decrease of accuracy. In such circumstance, the effectiveness of all the implemented methods has been distinctly weakened. However, the discriminant power of SMMFA is still remarkable since the recognition rate of SMMFA is much greater than those of other methods when more images are selected for training.

### 4.4 Experiments on GEMEP-FERA database

The Geneva multimodal emotion portrayals facial expression recognition and analysis (GEMEP-FERA) database is a subset of the GEMEP (Geneva multimodal emotion portrayals) corpus used as a database for the FERA (facial expression recognition and analysis) 2011 challenge [41]. FERA consists of sequences of 10 actors displaying different expressions. There are seven subjects in the training set and six subjects in the test set. Each sequence shows facial expressions of five emotion categories: anger, fear, joy, relief or sadness. We extract static frames from each sequence from both the training set and the test set, which results in 300 images including 67 fear, 57 sadness, 48 relief, 71 joy and 57 anger images. The extracted images are cropped and resized to the size of 32 × 32. Figure 8 shows samples of the cropped images. As we can see from Fig. 8, there are clear differences among the expression images in brightness, intensity of facial movements and head pose, which makes the recognition task more challenging.

We design similar experiments as we perform on JAFFE and CK+ databases. Let  $s \in \{4, 8, 12, 16, 20, 24\}$ . From each expression category,  $s$  images are randomly selected for training. For each  $s$ , the experiment is repeated 20 times to avoid accidental error. The best average

accuracy of each approach, with the standard deviation and the corresponding optimal dimension is given in Table 5.

Table 5 shows that the performance of all the seven methods is quite limited since their recognition rates are all below 65% even if 24 images from each expression class are selected for training. The main reason is that these expression images are clearly different in brightness, head pose and the intensity of facial movements. Naturally, all methods can enhance their performance using more training samples. As we

**Table 6** Results of statistical test

	$CD_{0,10}$	SMMFA	SLFDA	CMFA-SR	LFDA	LDA	MFA	LPP
Mean Rank	–	1.13	3.13	5.47	2.80	4.13	4.33	7.00
Friedman Test	1.45	–	2.00	4.33	1.67	3.00	3.20	5.87

can see from Table 5, SMMFA is better than the other algorithms on the GEMEP-FERA database.

#### 4.5 Statistical comparison

As described above, we perform experiments on six-group datasets of JAFFE, three-group datasets of CK+ and six-group datasets of GEMEP-FERA. Thus, there are a total of fifteen datasets. In order to compare SMMFA with other related methods more concretely, we perform the Friedman test over different datasets. The Friedman test is one kind of non-parametric statistical test which is used to determine whether there are differences across multiple datasets [42]. The critical difference (CD) is defined as follows:

$$CD_{\alpha} = q_{\alpha} \sqrt{\frac{j(j+1)}{6T}} \quad (14)$$

where  $\alpha$  is the threshold value which is set as 0.1 generally,  $j$  is the number of approaches in the experiments, and  $T$  is the number of datasets. Here,  $T = 15$ ,  $j = 7$  and  $q_{0.1} = 2.326$ , then we get the value of  $CD_{0.1}$ , that is, 1.45, by the calculation of (14).

Table 6 shows the results of the statistical test. The first row in Table 6 lists the mean rank of seven methods while the second row lists the Friedman results with respect to SMMFA. Since the Friedman results of LDA, MFA, CMFA-SR, LFDA, LPP and SLFDA are greater than the value of  $CD_{0.1}$ , the differences between them and SMMFA are significant, which means the performance of SMMFA is remarkably better than these six methods.

#### 5 Conclusion

In this paper, we propose a novel sparse subspace learning method for FER. First, the modified MFA is proposed to generate the original projection matrix, and then the linearized Bregman iteration is applied to obtain the sparse projection for SMMFA. Extensive experiments are conducted. On two toy datasets, we validate that SMMFA has the best separability among six compared methods. Experiments are also performed on three typical FER databases. On the JAFFE database, SMMFA is compared with or better than SLFDA, and much better than LPP, MFA, CMFA-SR, LDA and LFDA. On the CK+ database, the advantage of SMMFA is very obvious. On the GEMEP-FERA database, the sparse method SLFDA is distinctly weakened and SMMFA still achieves the highest recognition rate among all seven methods. All results in experiments show that SMMFA can effectively implement the task of dimension reduction and sparsification, obtaining a model with high degree of interpretability and generalization ability.

**Acknowledgements** This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61373093 and 61572339, by the Soochow Scholar Project, by the Six Talent Peak

Project of Jiangsu Province of China, and by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### References

1. Yuan C, Wu Q, Li P, et al (2018) Expression recognition algorithm based on the relative relationship of the facial landmarks. In: International Congress on Image & Signal Processing, Shanghai, China, pp 1–5
2. Liu X, Kumar BVKV, You J, et al (2017) Adaptive deep metric learning for identity-aware facial expression recognition. In: IEEE Conference on Computer Vision & Pattern Recognition Workshops, Honolulu, Hawaii, pp 522–531
3. Kabir MH, Salekin MS, Uddin MZ, Abdullah-al-Wadud M (2017) Facial expression recognition from depth video with patterns of oriented motion flow. IEEE ACCS 5(99):8880–8889
4. Vrigkas M, Nikou C, Kakadiaris IA (2016) Exploiting privileged information for facial expression recognition. In: International Conference on Biometrics, Halmstad, Sweden, pp 1–8
5. Wang Z, Ruan Q, An G (2016) Facial expression recognition using sparse local fisher discriminant analysis. Neurocomputing 174:756–766
6. Ren F, Huang Z (2015) Facial expression recognition based on AAM-SIFT and adaptive regional weighting. IEEE Trans Electr Electron Eng 10(6):713–722
7. Ekman P, Friesen W (1978) Facial action coding system: a technique for the measurement of facial action. Consulting Psychologists Press, Palo Alto
8. Amini R, Lisetti C, Ruiz G (2015) HapFACS 3.0: FACS-based facial expression generator for 3D speaking virtual characters. IEEE Trans Affect Comput 6(4):348–360
9. Hofmann J, Platt T, Ruch W (2017) Laughter and smiling in 16 positive emotions. IEEE Trans Affect Comput 8(4):495–507
10. Kamarol SKA, Jaward MH, Parkkinen J et al (2016) Spatiotemporal feature extraction for facial expression recognition. IET Image Process 10(7):534–541
11. Sun Y, Yu J (2017) Facial Expression Recognition by Fusing Gabor and Local Binary Pattern Features. In: International Conference on Multimedia modelling, Reykjavik, Iceland, pp 209–220
12. Puthenputhussery A, Liu Q, Liu C (2017) A sparse representation model using the complete marginal fisher analysis framework and its applications to visual recognition. IEEE Trans Multimedia 19(8):1757–1770
13. Zheng W, Zong Y, Zhou X, Xin M (2018) Cross-domain color facial expression recognition using transductive transfer subspace learning. IEEE Trans Affect Comput 9(1):21–37
14. Lin C, Long F, Zhan Y (2018) Facial expression recognition by learning spatiotemporal features with multi-layer independent subspace analysis. In: International Congress on Image & Signal Processing, Shanghai, China, pp 1–6
15. Nikitidis S, Tefas A, Pitas I (2013) Maximum margin discriminant projections for facial expression recognition. In: IEEE International Conference on Signal Processing, Marrakech, Morocco, pp 1–5
16. Jia J, Xu Y, Zhang S, et al (2016) The facial expression recognition method of random forest based on improved PCA extracting feature. In: IEEE International Conference on Signal Processing, Communications and Computing, Hong Kong, China, pp 1–5
17. Bouwmans T, Javed S, Zhang H, Lin Z, Otazo R (2018) On the applications of robust PCA in image and video processing. Proc IEEE 106(8):1427–1457
18. Imran MA, Miah MSU, Rahman H (2015) Face recognition using eigenfaces. Int J Comput Appl 118(5):12–16

19. Chao L, Ding J, Liu Z (2015) Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection. *Signal Process* 117(12):1–10
20. Chen SB, Wang J, Liu CY, Luo B (2017) Two-dimensional discriminant locality preserving projection based on  $\ell_1$ -norm maximization. *Pattern Recogn Lett* 87:147–154
21. Siddiqi MH, Ali R, Khan AM, Young-Tack Park, Sungyoung Lee (2015) Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. *IEEE Trans Image Process* 24(4):1386–1398
22. Tian C, Zhang Q, Sun G, et al (2016) Linear discriminant analysis representation and CRC representation for image classification. In: *IEEE International Conference on Computer & Communications*, Chengdu, China, pp 755–760
23. Shah JH, Sharif M, Yasmin M, Fernandes SL (2017) Facial expressions classification and false label reduction using LDA and threefold SVM. *Pattern Recognition Letters*. Available online 23 June 2017: <https://doi.org/10.1016/j.patrec.2017.06.021>
24. Sharma A, Paliwal KK (2015) Linear discriminant analysis for the small sample size problem: an overview. *Int J Mach Learn Cybern* 6(3):443–454
25. Yan S, Xu D, Zhang B, Zhang HJ, Yang Q, Lin S (2007) Graph embedding: a general framework for dimensionality reduction. *IEEE Trans Pattern Anal Mach Intell* 29(1):40–51
26. Lu GF, Zou J, Wang Y, Wang Z (2017)  $L_1$ -norm based null space discriminant analysis. *Multim Tools Appl* 76(14):15801–15816
27. Yin J, Jin Z (2012) From NLDA to LDA/GSVD: a modified NLDA algorithm. *Neural Comput & Applic* 21(7):1575–1583
28. Chu D, Liao LZ, Ng KP et al (2017) Incremental linear discriminant analysis: a fast algorithm and comparisons. *IEEE Trans Neural Networks Learn Syst* 26(11):2716–2735
29. Zhang L, Cobzas D, Wilman AH, Kong L (2018) Significant anatomy detection through sparse classification: a comparative study. *IEEE Trans Med Imaging* 37(1):128–137
30. Tibshirani R (1996) Regression shrinkage and selection via the LASSO: a retrospective. *J R Stat Soc* 58(1):267–288
31. Qiao T, Li W, Wu B (2014) A new algorithm based on linearized Bregman iteration with generalized inverse for compressed sensing. *Circuits Systems & Signal Processing* 33(5):1527–1539
32. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. *J Comput Graph Stat* 15(2):265–286
33. Qiao Z, Zhou L, Huang JZ (2009) Sparse linear discriminant analysis with applications to high dimensional low sample size data. *IAENG Int J Appl Math* 9(1):48–60
34. Chu D, Liao LZ, Ng MK, Zhang X (2013) Sparse canonical correlation analysis: new formulation and algorithm. *IEEE Trans Pattern Anal Mach Intell* 35(12):3050–3065
35. Jeribi A (2015) Spectral graph theory. In: *Spectral Theory and Applications of Linear Operators and Block Operator Matrices*, Springer, Cham, pp 413–439
36. Cai JF, Osher S, Shen Z (2009) Linearized Bregman iterations for compressed sensing. *Math Comput* 78(267):1515–1536
37. Huang B, Ma S, Goldfarb D (2013) Accelerated linearized Bregman method. *J Sci Comput* 54(2–3):428–453
38. Chen C, Xu G (2016) A new linearized split Bregman iterative algorithm for image reconstruction in sparse-view X-ray computed tomography. *Comput Math Appl* 71(8):1537–1559
39. Lyons M, Akamatsu S, Kamachi M, et al (1998) Coding facial expressions with Gabor wavelets. In: *Proceedings of the Third IEEE Conference on Face and Gesture Recognition*, Nara, Japan, pp 200–205
40. Lee K, Ho J, Kriegman D (2005) Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans Pattern Anal Mach Intell* 27(5):684–698
41. Valstar MF, Jiang B, Mehu M, et al (2011) The first facial expression recognition and analysis challenge. In: *IEEE International*

*Conference on Automatic Face & Gesture Recognition and Workshops*, pp 921–926

42. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 32(200):675–701



**Zhe Wang** was born in Nantong, Jiangsu province, China. He received the B.S. degree from the School of Computer Science and Technology, Soochow University, Suzhou, China, in 2018. He is currently working in Huawei Technologies Co., Ltd. His research interests include machine learning, data mining, and pattern recognition.



**Li Zhang** received the B.S. degree in 1997 and the Ph.D. degree in 2002 in electronic engineering from Xidian University, Xi'an, China. Now she is a full professor with the School of Computer Science and Technology, Soochow University, Suzhou, China. She was a postdoctor at the Institute of Automation, Shanghai Jiao Tong University, Shanghai, China, from 2003 to 2005. She worked as an associate professor at the Institute of Intelligent Information

Processing, Xidian University, Xi'an, China, from 2005 to 2010. She was a visiting professor at Yuan Ze University, Taiwan, from February to May 2010. She has authored/co-authored more than 100 technical papers published in journals and conferences. Her research interests have been in the areas of machine learning, pattern recognition, neural networks and intelligent information processing.



**Bangjun Wang** received the M.S. degree in information management from the Wuhan University of Technology, Wuhan, in 2004. From 2004, he worked in Soochow University. He is currently working toward the Ph.D. degree in Beijing Jiaotong University, Beijing, China. His research interests include machine learning and Layered learning.