Sparse Tensor Discriminant Analysis

Zhihui Lai, Yong Xu, *Member, IEEE*, Jian Yang, Jinhui Tang, and David Zhang, *Fellow, IEEE*

***Abstract*— The classical linear discriminant analysis has undergone great development and has recently been extended to different cases. In this paper, a novel discriminant subspace learning method called sparse tensor discriminant analysis (STDA) is proposed, which further extends the recently presented multilinear discriminant analysis to a sparse case. Through introducing the L_1 and L_2 norms into the objective function of STDA, we can obtain multiple interrelated sparse discriminant subspaces for feature extraction. As there are no closed-form solutions, k-mode optimization technique and the L_1 norm sparse regression are combined to iteratively learn the optimal sparse discriminant subspace along different modes of the tensors. Moreover, each non-zero element in each subspace is selected from the most important variables/factors, and thus STDA has the potential to perform better than other discriminant subspace methods. Extensive experiments on face databases (Yale, FERET, and CMU PIE face databases) and the Weizmann action database show that the proposed STDA algorithm demonstrates the most competitive performance against the compared tensor-based methods, particularly in small sample sizes.**

***Index Terms*— Linear discriminant analysis, feature extraction, sparse projections, face recognition.**

I. Introduction

As the classical supervised subspace learning method, Linear Discriminant Analysis (LDA) [1]–[3] has undergone continuous development for several decades. The traditional LDA algorithms treat the input image object as a high-dimensional vector (1D vector) by concatenating the rows or columns of the images. The image-to-vector transform procedure usually causes the so called "curse of dimensionality" [4] and the loss of some useful structural information embedded in the original images.

However, some recent works show that to operate directly on the image matrices for feature extraction can obtain better classification performance. Yang et al. extended the classical Principal Component Analysis (PCA) [5]–[7] to image matrices based representation and proposed the well-known Two-Dimensional PCA (2DPCA) [8]. With the similar idea of 2DPCA, researchers proposed Two Dimensional LDA (2DLDA) [9], [10] for extracting features from the image matrices. Since the size of the covariance matrices constructed by the 2D methods is significantly smaller than the ones in the classical PCA and LDA, they are more efficient and effective than the corresponding traditional methods in small sample size problems. A common disadvantage of these 2D methods is that a single projection matrix is learned for dimensionality reduction from one side of the image matrix, and thus more coefficients are needed to represent the image matrix. In order to solve this problem, Generalized Low Rank Approximations of Matrices (GLRAM) [11] and bidirectional LDA [12] were also developed for image feature extraction. For more details of the 2D based methods, the readers are referred to see [13]–[16].

High dimensional vector is the first order tensor, and one projection matrix can be obtained. The image matrix is the tensor of order 2, and thus we can obtain the bidirectional projection matrices for feature extraction. A natural way is to compute n projection matrices for nth-order tensor feature extraction. In recent years, there was great interest in high-order tensor feature extraction, and higher order tensor decomposition [17]–[19] has become an important technique in computer vision and pattern recognition [20]–[23]. More recently, Concurrent Subspaces Analysis [24], Multilinear PCA (MPCA) [25] and its uncorrelated variation [26] were proposed for face and gait recognition tasks. In order to enhance the performance of the tensor-based method for classification, Multilinear Discriminant Analysis (MDA) [27], which generalized traditional LDA to tensor based LDA, was also developed for face recognition. By considering the uncorrelated property between the features, Lu et al. [28] proposed uncorrelated multilinear discriminant analysis for tensor object recognition. Unfortunately, as stated in [16], the ratio-base multilinear discriminant analysis methods do not converge and appear to be extremely sensitive to parameter settings. Therefore, Tao et al. proposed General Tensor Discriminant Analysis (GTDA) [29] for gait recognition using the Differential Scatter Discriminant Criterion (DSDC) [30], and Hu et al. proposed the Tensor Maximum Margin Criterion (TMMC) [31] for object recognition using the Maximum

Manuscript received June 14, 2012; revised October 5, 2012, February 2, 2013, and April 12, 2013; accepted May 13, 2013. Date of publication May 22, 2013; date of current version August 30, 2013. This work was supported in part by the Natural Science Foundation of China under Grants 61203376, 60973098, 61005005, 61071179, 61125305, and 61173104, the Central Fund from the Hong Kong Polytechnic University, the Fundamental Research Funds for the Central Universities under Grant HIT.NSRIF. 2009130, the China Postdoctoral Science Foundation under Project 2012M510958, the Guangdong Natural Science Foundation under Project S2012040007289, Shenzhen Council for Scientific and Technological Innovation under Grant JCYJ20120613134843060, and the Program for New Century Excellent Talents in University under Grant NCET-12-0632, and the Natural Science Foundation of Jiangsu Province under Grants BK2012033 and BK2011700. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rebecca Willett.

Z. Lai is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: lai_zhi_hui@163.com).

Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: laterfall2@yahoo.com.cn).

J. Yang and J. Tang are with the School of Computer Science, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China (e-mail: csjyang@mail.njust.edu.cn; tangjh1981@acm.org).

D. Zhang is with the Biometrics Research Centre, Department of Computing, Hong Kong Polytechnic University, Kowloon 999077, Hong Kong (e-mail: csdzhang@comp.polyu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2013.2264678

Margin Criterion (MMC) [32]. Similarly, rank-one decomposition method was also developed for tensor learning [33]. The readers are referred to a survey [34] of multilinear subspace learning for more information.

The tensor data contain large quantities of information redundancy and thus not all the features/variables are important for feature extraction and classification [25], [26], [35]. Therefore, a method that can remove the redundancy information or "filter" out the unimportant features in feature extraction step is much appreciated. It can be found from the literature that using the sparse representation/learning methods to perform the feature selection on the projection vectors/matrices can achieve this end. The sparse representation methods introduce the L_1 norm, by which some representation coefficients can be shrunk to be zero, to achieve the goal of sparse feature selection and classifier design. For example, the reconstruction errors can be used in sparse representation classifier for robust face recognition [36]. In [37], Tibshirani proposed the Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection and obtained better performance than the ordinary least squares regression. By combining both L_1 and L_2 norms penalty, the Elastic Net [38] proposed by Zou and Hastie generalizes the LASSO and overcome some potential drawbacks of LASSO by using Least Angle Regression (LARS) [39] to obtain the optimal solution path. Sparse Principal Component Analysis (SPCA) [40] reformulates the PCA as regression-type optimization and realizes the goal of spare feature selection and dimensionality reduction with interpretabilities. In SPCA, the most important/contributive variables for scatter distribution are selected to form the principal components. However, SPCA is an unsupervised method and its performance in discriminant ability is relatively poor. In order to enhance the classification performance, Sparse Discriminant Analysis (SDA) [41] and Sparse Linear Discriminant Analysis (SLDA) [42], which use the sparse regression for discriminant features/variables selection, were proposed to learn a sparse discriminant subspace for feature extraction and classification in biological and medical data analysis. These sparse regression based methods have been proved to be effective for classification and prediction.

Although there are some related works on sparse tensor learning such as tensor sparse coding [43] and sparse non-negative tensor factorization [44], [45], until now, high order tensor data analysis for feature extraction with sparse manner has not been widely investigated and how to extend the discriminant analysis algorithms with sparseness to the high order tensor learning is unsolved. In this paper, motivated by SPCA, SDA and the tensor-based GTDA algorithms, we proposed Sparse Tensor Discriminant Analysis (STDA) for feature extraction and classification. Our starting point is to introduce the L_1 and L_2 norms penalty on the projection vectors/matrices and use sparse regression method to select the most discriminant features/variables to form the projections. By doing this, a set of interrelated sparse projection vectors/matrices formed by the important discriminant variables are obtained, and thus the discriminant ability of the learned subspaces are potentially more powerful than other methods' on tensor data.

STDA has two significant properties that the previous tensor-based feature extraction methods do not hold. Firstly, different from the previous tensor-based methods such as the MPCA, MDA and GTDA, all the projection matrices of STDA derived from different modes are sparse. Secondly, the optimal multi-linear sparse projections of STDA are obtained by iterating the Elastic Net regression and Singular Value Decomposition (SVD) instead of solving the eigenequations as in MPCA, MDA and GTDA .

The main contributions of this paper are as follows: First, we propose the sparse difference (tensor) scatter criterion for sparse subspace learning. With the proposed criterion, all the MMC-based or DSDC-based methods such as those in [29]–[32], [46]–[48] can be similarly extended to sparse cases. Thus, the proposed method can be used as a unified framework for learning the sparse multi-linear projections. Second, the relationships between STDA and other algorithms, i.e. MMC, TMMC, GTDA etc., are theoretically analyzed. Third, STDA outperforms the MMC-based or DSDC-based methods and their higher order (second and third orders) extensions in classification accuracy.

The rest of the paper is organized as follows. In Section II, STDA algorithm and related analyses are presented. In Section III, the theoretical analysis is performed for exploring the relationships between STDA and the previous methods. Experimental results to evaluate the STDA algorithm are shown in Section IV, and the conclusions are given in Section V.

II. Sparse Tensor Discriminant Analysis

In this section, we briefly review some basic multilinear notations, definitions and operations at first and then present the STDA algorithm.

A. Preparations

In this paper, if there are no special instructions, lowercase and uppercase italic letters, i.e. $i, j, m, k, \alpha, \beta, N$ etc., denote scalars, bold lowercase letters, i.e. $\mathbf{a}, \mathbf{b}, \mathbf{u}$ etc., denote vectors, and bold uppercase letters, i.e. $\mathbf{A}, \mathbf{B}, \mathbf{S}, \mathbf{\Phi}$ etc., denote the matrices, and the Lucida calligraphy Italic letters, i.e. $\mathcal{X}, \mathcal{Y}$ denote the tensors.

Assume that the training samples are represented as the nth-order tensor $\{\mathcal{X}_i \in R^{m_1 \times m_2 \times \cdots \times m_n}, i = 1, 2, \ldots, N\}$, where N denotes the total number of the training samples. Moreover, let N_c denote the total number of classes and N_{c_i} denote the number of all the samples in the ith class.

Definition 1: The inner product of two tensors $\mathcal{X}, \mathcal{Y} \in R^{m_1 \times m_2 \times \cdots \times m_n}$ is defined as $< \mathcal{X}, \mathcal{Y} >= \sum_{i_1,=1,\ldots,i_n=1}^{m_1 \times m_2 \times \cdots m_n} \mathcal{X}_{i_1,\ldots,i_n} \mathcal{Y}_{i_1,\ldots,i_n}$. The norm of a tensor is defined as $||\mathcal{X}|| = \sqrt{< \mathcal{X}, \mathcal{X} >}$. The tensor distance between two tensors $\mathcal{X}$ and $\mathcal{Y}$ is defined as $D(\mathcal{X}, \mathcal{Y}) = ||\mathcal{X} - \mathcal{Y}||$.

Definition 2: The mode-k flattening of the nth-order tensor $\mathcal{X} \in R^{m_1 \times m_2 \times \cdots \times m_n}$ into a matrix $X^{(k)} \in R^{m_k \times \prod_{i \neq k} m_i}$, i.e. $\mathbf{X}^{(k)} \Longleftarrow_k \mathcal{X}$, is defined as $\mathbf{X}^{(k)}_{i_k, j} = \mathcal{X}_{i_1, i_2, \ldots, i_n}$, where $j = 1 + \sum_{l=1, l \neq k}^{n} (i_l - 1) \prod_{o=l+1, o \neq k}^{n} m_o$.

Definition 3: The mode-k product of tensor $\mathcal{X}$ with matrix $\mathbf{U} \in R^{m'_k \times m_k}$ is defined as $\mathcal{Y} = \mathcal{X} \times_k \mathbf{U}$, where

$\mathcal{Y}_{i_1},\ldots,i_{k-1},i,i_{k+1},\ldots,i_n = \sum_{j=1}^{m'_k} \mathcal{X}_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_n}\mathbf{U}_{i,j}$ $(j=1,\ldots,m'_k)$ and $\mathbf{U}_{i,j}$ denotes the element in the matrix $\mathbf{U}$ of coordinate (i,j).

Definition 4: The mode-k within-class tensor scatter matrix $\mathbf{S}_w^{(k)}$ is defined as

$$\mathbf{S}_w^{(k)} = \sum_{j=1}^{N_c}\sum_{i=1}^{N_{c_i}} \left(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}}_j^{(k)}\right)\left(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}}_j^{(k)}\right)^T$$

where $\bar{\mathbf{X}}_j^{(k)}$ denotes the mean value of the mode-k flattening of the tensor samples in the jth class.

Definition 5: The mode-k between-class tensor scatter matrix $\mathbf{S}_B^{(k)}$ is defined as

$$\mathbf{S}_B^{(k)} = \sum_{i=1}^{N_c} N_{C_1}\left(\bar{\mathbf{X}}_i^{(k)} - \bar{\mathbf{X}}^{(k)}\right)\left(\bar{\mathbf{X}}_i^{(k)} - \bar{\mathbf{X}}^{(k)}\right)^T$$

where $\bar{\mathbf{X}}^{(k)}$ denotes the mean value of the mode-k flattening of the tensor samples of all the training samples.

With the above preparations, we can directly present the objective function of STDA in the following section.

B. The Objective Function of STDA

The purpose of the STDA is to obtain a set of sparse projection matrix $\{\mathbf{U}_i \in R^{m_i\times d_i}, d_i \le m_i, i=1,2,\ldots,n\}$ that map the original high-order tensor data into a low-order tensor space:

$$\mathcal{Y}_i = \mathcal{X}_i \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \cdots \times_n \mathbf{U}_n^T. \tag{1}$$

The objective function of STDA is to minimize the tensor discriminant function of the L_1 and L_2 norms penalty optimization problem with a set of constraints:

$$\mathbf{U}_k^* = \arg\min \operatorname{tr}(\mathbf{U}_k^T(\mathbf{S}_w^{(k)} - \mu\mathbf{S}_B^{(k)})\mathbf{U}_k) + \boldsymbol{\alpha}_k||\mathbf{U}_k||^2 + \sum_j \beta_{\mathrm{kj}}|\mathbf{u}_{\mathrm{kj}}| \tag{2}$$

$$subject\ to\ \mathbf{U}_k^T\mathbf{U}_k = \mathbf{I}_k \tag{3}$$

where $k=1,2,\ldots,n$, μ is a suitable constant set by the user and $\mathbf{u}_{\mathrm{kj}}$ is the j-th column of $\mathbf{U}_k$, $||\cdot||$, and $|\cdot|$ denote the L_2 and L_1 norm, respectively. $\boldsymbol{\alpha}_k s$ are the constants set by the user and β_{kj} s are the coefficients of L_1 norm which can be optimally determined by the Elastic Net. Similar techniques exist in [40]–[42].

The above optimization problem uses the criterion negative to MMC (or TMMC) but with the additional L_1 and L_2 norms for obtaining the sparse projections. Thus, we refer to it as the sparse difference (tensor) scatter criterion. Though it is formally different from MMC criterion, the intrinsic geometric meanings of these two criterions are very similar. To minimize the first term $\operatorname{tr}(\mathbf{U}_k^T(\mathbf{S}_w^{(k)} - \mu\mathbf{S}_B^{(k)})\mathbf{U}_k)$ also means that the projections should enable the within-class scatter and between-class scatter to be minimized and maximized, respectively. If projections $\mathbf{U}_k$ can do so, they will be helpful for discrimination.

To the best of our knowledge, there exist no closed-form solutions for such complex objective function. Fortunately, the optimization problem can be converted to a problem to independently find n subspaces $\mathbf{U}_k (k=1,2,\ldots,n)$ that can minimize the scatter value of the mode-k flattening of the nth-order tensors with L_1 and L_2 norms penalty. We can further obtain the approximate sparse solutions by rewriting the optimization problem as a set of independent SPCA criterions. In the following sections, we first explore the effective discriminant projections, and then we focus on the mode-k flattening of the nth-order tensors to obtain the sparse discriminant projections. The key problem is to convert the sparse optimization model into the models that are easy to solve.

Differing from the previous tensor discriminant analysis methods in which each subspace is obtained by performing SVD or eigen decomposition, STDA uses the regression method with L_1 and L_2 norms penalty to obtain each sparse projections/subspace. In the following two sections, the sparse difference scatter criterion for discrimination is first analyzed, and then the effective sparse discriminant projections is obtained by combining the L_1 and L_2 norms penalty for regression. At last, the local optimal solutions can be obtained by alternative iterations.

C. Analysis on the Optimization Problem

At first, we analyze the first part of the sparse criterion in (2). Since the optimization problem $\min\operatorname{tr}(\mathbf{U}_k^T(\mathbf{S}_w^{(k)} - \mu\mathbf{S}^{(k)}{}_B)\mathbf{U}_k)$ is exactly the negative of $\max\operatorname{tr}(\mathbf{U}_k^T(\mathbf{S}_B^{(k)} - \mu\mathbf{S}_w^{(k)})\mathbf{U}_k)$ with the orthogonal constraint of $\mathbf{U}_k^T\mathbf{U}_k = \mathrm{I}_k$, the eigenvectors (or singular value vectors) corresponding to the first d_i minimal eigenvalues of $\mathbf{S}_w^{(k)} - \mu\mathbf{S}_B^{(k)}$ are exactly the eigenvectors corresponding to the first d_i maximal eigenvalues of $\mathbf{S}_B^{(k)} - \mu\mathbf{S}_w^{(k)}$. Therefore, in the following sections, we only focus on the minimum optimization problem to obtain the optimal sparse projection for dimensionality reduction. The reason for taking the minimum problem into consideration is that the sparse projections are computed by the Elastic Net which is also the minimum optimization problem and thus the minimum optimization problem $\min tr(\mathbf{U}_k^T(\mathbf{S}_W^{(k)} - \mu\mathbf{S}_B^{(k)})\mathbf{U}_k)$ and sparse regression problem can be integrated together.

Denote the SVD of $\mathbf{S}_W^{(k)} - \mu\mathbf{S}_B^{(k)}$ as follows:

$$\mathbf{S}_W^{(k)} - \mu\mathbf{S}_B^{(k)} = \boldsymbol{\Phi}_k\boldsymbol{\Lambda}_k\boldsymbol{\Phi}_k^T \tag{4}$$

where $\boldsymbol{\Phi}_k$ and $\boldsymbol{\Lambda}_k = \operatorname{diag}[\boldsymbol{\Lambda}_k^1, \boldsymbol{\Lambda}_k^2, \ldots, \boldsymbol{\Lambda}_k^{m_k}]$ are the left SVD matrix and the corresponding singular values sorted in ascending order.

Let $\mathbf{S}^{(k)} = \mathbf{S}_W^{(k)} - \mu\mathbf{S}_B^{(k)} = \boldsymbol{\Phi}_k\boldsymbol{\Lambda}_k\boldsymbol{\Phi}_k^T$, then $\mathbf{S}^{(k)}$ can be rewritten as

$$\mathbf{S}^{(k)} = \boldsymbol{\Phi}_k\boldsymbol{\Lambda}_k\boldsymbol{\Phi}_k^T = \boldsymbol{\Phi}_k\sqrt{\boldsymbol{\Lambda}_k}\boldsymbol{\Phi}_k^T(\boldsymbol{\Phi}_k\sqrt{\boldsymbol{\Lambda}_k}\boldsymbol{\Phi}_k^T)^T = \hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^T \tag{5}$$

where $\hat{\mathbf{X}}_k = \boldsymbol{\Phi}_k\sqrt{\boldsymbol{\Lambda}_k}\boldsymbol{\Phi}_k^T$ and $\sqrt{\boldsymbol{\Lambda}_k}$ is defined as $(\sqrt{\boldsymbol{\Lambda}_k})_i = \left\{\begin{matrix}\sqrt{\boldsymbol{\Lambda}_k^i}, & \text{if } \boldsymbol{\Lambda}_k^i \ge 0\\ -\sqrt{-\boldsymbol{\Lambda}_k^i}, & \text{if } \boldsymbol{\Lambda}_k^i < 0\end{matrix}\right\}$, where $(\cdot)_i$ denotes the i-th element of the diagonal matrix. Then the first part of optimization problem (2) can be equivalently represented as the following

optimization problem to compute the projection matrix $\mathbf{U}_k$

$$\min tr(\mathbf{U}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k) \quad \textit{subject to } \mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_k \tag{6}$$

It can be checked that $\mathbf{U}_k = \mathbf{\Phi}_k$ is exactly the solution of the above optimization problem (6). The optimal solution can be obtained by performing SVD of $\hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T$ or solving the standard eigen-function $\hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k = \mathbf{\Lambda}_k \mathbf{U}_k$. However, the SVD or eigen-decomposition can not obtain the optimal sparse solution. Thus, we will introduce the L_1 and L_2 norms penalty into the optimization procedures to compute the sparse solution, which is presented in the sections D and E.

D. Regression Analysis on the Discriminant Projections

Since directly solving the eigen-function cannot provide the sparse solutions, one option is to convert the optimization problem into regression forms such that the previous sparse regression methods can be used to compute the sparse projection. For this end, the relationship between the optimization problem of eigen-function (6) and the regression problem should be firstly revealed. Let us take the following optimization problem into consideration:

$$\mathbf{U}_k^* = \arg\min_{\mathbf{U}_k} ||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{U}_k^T \hat{\mathbf{X}}_k||^2 + \boldsymbol{\alpha}_k ||\mathbf{U}_k||^2. \tag{7}$$

The relationship between optimization problem (6) and (7) can be revealed by the following theorem.

Theorem 1: The solution space of (6) is equivalent to the one of (7). Moreover, let $\mathbf{u}_{kj}^*$ be the j-th column of $\mathbf{U}_k^*$, then $\mathbf{u}_{kj}^* \propto \varphi_{kj}$ where φ_{kj} denotes the column vector in $\mathbf{\Phi}_k$.

Proof: From (7) we have

$$\begin{aligned} &||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{U}_k^T \hat{\mathbf{X}}_k||^2 + \boldsymbol{\alpha}_k ||\mathbf{U}_k||^2 \\ &= tr(\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{\Phi}_k - 2\mathbf{U}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{\Phi}_k + \mathbf{U}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k + \boldsymbol{\alpha} \mathbf{U}_k^T \mathbf{U}_k) \end{aligned}$$

Taking the derivation with respect to $\mathbf{U}_k$ be 0, we can represent the optimal solution of the above (regression) problem as

$$\begin{aligned} \mathbf{U}_k^* &= (\hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T + \boldsymbol{\alpha}_k \mathbf{I})^{-1} \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{\Phi}_k \\ &= (\mathbf{\Phi}_k \sqrt{\mathbf{\Lambda}_k} \mathbf{\Phi}_k^T \mathbf{\Phi}_k \sqrt{\mathbf{\Lambda}_k} \mathbf{\Phi}_k^T + \boldsymbol{\alpha}_k \mathbf{I})^{-1} \\ &\quad \times \mathbf{\Phi}_k \sqrt{\mathbf{\Lambda}_k} \mathbf{\Phi}_k^T \mathbf{\Phi}_k \sqrt{\mathbf{\Lambda}_k} \mathbf{\Phi}_k^T \mathbf{\Phi}_k \\ &= (\mathbf{\Phi}_k (\mathbf{\Lambda}_k + \boldsymbol{\alpha}_k \mathbf{I}) \mathbf{\Phi}_k^T)^{-1} \mathbf{\Phi}_k \mathbf{\Lambda}_k \\ &= \mathbf{\Phi}_k \frac{\mathbf{\Lambda}_k}{\mathbf{\Lambda}_k + \boldsymbol{\alpha}_k \mathbf{I}} \end{aligned}$$

From the above equation, we can see that (6) and (7) have the same solution space, i.e. $\mathbf{u}_{kj}^* \propto \varphi_{kj}$. ■

Corollary 1: In regression problem (7), if $\boldsymbol{\alpha}_k \to 0$, then $\mathbf{U}_k^* \to \mathbf{\Phi}_k$. If $\boldsymbol{\alpha}_k = 0$, then $\mathbf{U}_k^* = \mathbf{\Phi}_k$.

Furthermore, we have the following theorem which generalizes the representation of the problem in (6) or (7) when $\boldsymbol{\alpha}_k = 0$.

Theorem 2: Suppose $\mathbf{U}_k$ is an unknown $m_i \times m_i$ matrix with unit column vector. For any given $m_i \times m_i$ matrix $\mathbf{A}_k$ satisfying $\mathbf{A}_k^T \mathbf{A}_k = \mathbf{I}_k$, the following optimization problem

$$\mathbf{U}_k^* = \arg\min_{\mathbf{U}_k, \mathbf{A}_k} ||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{A}_k \mathbf{U}_k^T \hat{\mathbf{X}}_k||^2 \tag{8}$$

has the same solution space as the optimization problem (6).

Proof: Since $\mathbf{A}_k^T \mathbf{A}_k = \mathbf{I}_k$ and $\mathbf{\Phi}_k^T \mathbf{\Phi}_k = \mathbf{I}_k$

$$\begin{aligned} &||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{A}_k \mathbf{U}_k^T \hat{\mathbf{X}}_k||^2 \\ &= tr(\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{\Phi}_k + \mathbf{A}_k \mathbf{U}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k \mathbf{A}_k^T - 2\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k \mathbf{A}_k^T) \\ &= tr(\hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T + \mathbf{U}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k - 2\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k \mathbf{A}_k^T) \end{aligned}$$

The last term can be rewritten as

$$\begin{aligned} tr(\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k \mathbf{A}_k^T) &= tr(\mathbf{A}_k^T \mathbf{\Phi}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k) \\ &= tr(\mathbf{A}_k^T \mathbf{\Phi}_k^T \mathbf{\Phi}_k \sqrt{\mathbf{\Lambda}_k} \mathbf{\Phi}_k^T \mathbf{\Phi}_k \sqrt{\mathbf{\Lambda}_k} \mathbf{\Phi}_k^T \mathbf{U}_k) \\ &= tr(\mathbf{A}_k^T \mathbf{\Lambda}_k \mathbf{\Phi}_k^T \mathbf{U}_k) \end{aligned}$$

It can be checked that if and only if $\mathbf{U}_k \equiv \mathbf{\Phi}_k \delta_k$, where δ_k is the diagonal matrix with 1 or -1, the above equation achieves the maximum $tr(\mathbf{A}_k^T \mathbf{\Lambda}_k \delta_k)$ and thus (8) achieve its minimum $2tr(\hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T - \mathbf{A}_k^T \mathbf{\Lambda}_k \delta_k)$. Therefore, optimization problem (8) has the same solution space as the optimization problem (6). ■

The above two theorems represent optimal problem (6) into a regression problem, which provide a tractable method for us to convert the optimization problem into previous regression framework with the L_1 and L_2 norms penalty. To this end, we should further bridge the connections between the problem in (7) or (8) with a more general form, which relaxes Theorem 2 by converting the orthogonal constraint on $\mathbf{A}_k$ to a direct additional constraint. Thus, the following result is obtained:

Theorem 3: The optimal solution space of (8) is the same as the following optimization problem:

$$\mathbf{U}_k^* = \arg\min ||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{A}_k \mathbf{U}_k^T \hat{\mathbf{X}}_k||^2 + \boldsymbol{\alpha}_k ||\mathbf{U}_k||^2 \quad \textit{subject to } \mathbf{A}_k^T \mathbf{A}_k = \mathbf{I}_k \tag{9}$$

That is $\mathbf{u}_{kj}^* \propto \varphi_{kj}$ for all $j = 1, 2, \ldots, d_k$.

Proof: The proof is presented in Appendix. ■

Corollary 2: The optimization problem (9) has the same solution space as (7) or (8), i.e. $span(\mathbf{U}_k) = span(\mathbf{\Phi}_k)$.

Thus, Theorem 3 reveals the relationships among the optimization problems (6), (7) and the constrained regression problem (9). The above analyses show that the constrained regression solution space is the same to the solution of (6).

At last, with the above preparations, the original optimization problem (2-3) can be converted to the following optimization problem to obtain the sparse discriminant projections of STDA on the mode-k flattening of the tensor data, in which the L_1 norm is added to (9) to form a new optimization problem:

$$\mathbf{U}_k^* = \arg\min ||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{A}_k \mathbf{U}_k^T \hat{\mathbf{X}}_k||^2 + \boldsymbol{\alpha}_k ||\mathbf{U}_k||^2 + \sum_j \beta_{kj} |\mathbf{u}_{kj}| \quad \textit{subject to } \mathbf{A}_k^T \mathbf{A}_k = \mathbf{I}_k \tag{10}$$

Therefore, optimization problem (10) provides the equivalent or approximate sparse solutions of the original problem (2–3).

E. STDA Algorithm and Its Related Analysis

It can be found that the above optimization problem (10) is exactly the modified version of SPCA criterion in [40].

TABLE I
STDA ALGORITHM

Input: Tensor samples $\{\mathcal{X}_i \in R^{m_1 \times m_2 \times \cdots \times m_n}, i = 1, 2, ..., N\}$, the numbers of iterations $T_{\max}$ and T_{EN}, the dimensions $d_i (\leq m_i), i = 1, 2, ..., n$
Output: Low-dimensional features $\mathcal{Y}_i (i = 1, 2, ..., N)$

Step 1: Center the training input samples.
Step 2: Initialize $\mathbf{U}_k^0|_{k=1}^n$ as arbitrary columnly-orthogonal matrices.
Step 3: For $t = 1{:}T_{\max}$ do
  For $k = 1{:}n$ do
    *Compute $\mathcal{X}_i^k{:}\mathcal{X}_i^k = \mathcal{X}_i \times_1 \mathbf{U}_1^{t-1T} \cdots \times_{k-1} \mathbf{U}_{k-1}^{t-1T} \times_{k+1} \mathbf{U}_{k+1}^{t-1T} \cdots \times_n \mathbf{U}_n^{t-1T} (i = 1, 2, \ldots, N)$
    *Perform the mode-k flattening of the nth-order tensors $\mathcal{X}_i^k$ to matrices: $X^{(k)} \Longleftarrow_k \mathcal{X}_i^k$
    *Compute the scatter matrices $S_B^{(k)}$ and $\mathbf{S}_W^{(k)}$ in definition 4 and 5.
    *Perform SVD on $\mathbf{S}_W^{(k)} - \mu \mathbf{S}_B^{(k)}$ to obtain $\hat{\mathbf{X}}_k$ and $\mathbf{\Phi}_k$
    *Initialize $\mathbf{A}_k$ as arbitrary columnly-orthogonal matrix
    *For $j = 1{:}T_{EN}$ do
      -Solve the Elastic Net problem: $\mathbf{U}_k^{t*} = \arg\min ||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{A}_k \mathbf{U}_k^{t-1T} \hat{\mathbf{X}}_k||^2 + \boldsymbol{\alpha}_k ||\mathbf{U}_k^{t-1}||^2 + \sum_j \beta_{kj} |\mathbf{u}_{kj}^{t-1}|$
      -Do SVD of $\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T U_k = \bar{\mathbf{U}}_k \bar{\mathbf{D}}_k \bar{\mathbf{V}}_k^T$, and update $\mathbf{A}_k = \bar{\mathbf{U}}_k \bar{\mathbf{V}}_k^T$.
    *Normalize $\mathbf{U}_k^{t*}$, i.e. let $\mathbf{U}_k^t(:, S) = \mathbf{U}_k^{t*}(:, S)/||\mathbf{U}_k^{t*}(:, S)||, S = 1{:}d_k$
    *If $t \geq 2$ and $\sum_k tr(\mathbf{U}_k^{tT} \mathbf{U}_k^t - \mathbf{I}_k) < n\varepsilon$, break.
Step 4: Project the tensor samples into the low-dimensional tensor subspace $\mathcal{Y}_i = \mathcal{X}_i \times_1 \mathbf{U}_1^{tT} \times_2 \mathbf{U}_2^{tT} \cdots \times_n \mathbf{U}_n^{tT}$

Thus, the algorithm procedures similar to SPCA can be used to compute the optimal sparse discriminant projections. The algorithm details of STDA are stated in Table I.

Computational Complexity: For simplicity, we assume that $m_1 = m_2 = \cdots = m_n = m$ and the total number of training samples N is comparable in magnitude to the feature dimension m^n. The complexity of MMC-based methods is $O(m^{3n})$. The main complexity of STDA is $O(tnT_{EN}m^3)$, where T_{EN} is the iteration number in the Elastic Net. Although many loops are required for STDA for optimization, it is still computationally more efficient than the high-dimensional vector based methods since the iteration numbers are usually small. In addition, computing the sparse discriminant projections is only needed in the training phase, therefore it can be done offline and the computational cost is acceptable.

Convergence of STDA: STDA can also converge very fast as TMMC. For the convergence of STDA, we have the following theorem.

Theorem 4: The iterative procedures of STDA presented in Table I will converge to a local optimum.

Proof: We need to prove that the objective function of STDA is non-increasing and has a lower bound (at least bigger than a constant $c > 0$). The original objective function of STDA in each iteration step can be rewritten as follow:

$$J(\mathbf{U}_1^t, \mathbf{U}_2^t, \ldots, \mathbf{U}_n^t) = tr(\mathbf{U}_k^{tT}(\mathbf{S}_W^{(k)} - \mu \mathbf{S}_B^{(k)})\mathbf{U}_k^t) + \boldsymbol{\alpha}_k ||\mathbf{U}_k^t||^2 + \sum_j \beta_{kj} |\mathbf{u}_{kj}^t|$$

where $\mathbf{U}_k^t$ denotes the tth iteration of $\mathbf{U}_k$ and $\mathbf{u}_{kj}^t$ is the column vector in $\mathbf{U}_k^t$. From the inner loop of the iteration by using the Elastic Net and SVD in STDA algorithm, we know that for each mode-k $\mathbf{U}_k^{t+1}$ makes the objective function achieve a local minimum according to theorems 1 and 3. Therefore, in the outer loop, we have

$$\begin{aligned} J(\mathbf{U}_1^t, \mathbf{U}_2^t, \ldots, \mathbf{U}_n^t) &\geq J(\mathbf{U}_1^{t+1}, \mathbf{U}_2^t, \ldots, \mathbf{U}_n^t) \\ &\geq J(\mathbf{U}_1^{t+1}, \mathbf{U}_2^{t+1}, \ldots, \mathbf{U}_n^t) \geq \cdots \\ &\geq J(\mathbf{U}_1^{t+1}, \mathbf{U}_2^{t+1}, \ldots, \mathbf{U}_n^{t+1} \geq c > 0 \end{aligned}$$

Therefore, the objective function will converge to a local optimum. ■

III. RELATION TO PREVIOUS WORKS

In this section, we discuss the relation between the proposed STDA algorithm and some previous dimensionality reduction methods i.e. PCA, SPCA, 2DPCA, CSA, MPCA, MMC, TMMC and GTDA. It is shown that these methods are the special cases of STDA. Therefore, STDA is a more general framework on data analysis.

A. Connection to PCA-Based Methods

At first, we reveal the relationship between STDA and MPCA. And then, it is natural to extend this relationship to the other special forms of MPCA.

The objective function of MPCA is to maximize the total scatter of the mode-k unfolding of the tensors. Its optimization problem can be represented as:

$$\begin{aligned} \mathbf{U}_k^* &= \arg\max tr(\mathbf{U}_k^T \mathbf{S}_T^{(k)} \mathbf{U}_k) \\ &s.t. \quad \mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_k \end{aligned}$$

Since $\mathbf{S}_T^{(k)} = \mathbf{S}_W^{(k)} + \mathbf{S}_B^{(k)}$, we have the following theorem.

Theorem 5: The projections of MPCA are the same as the projections of STDA when $\mu = -1$, $\mathbf{A}_k = \mathbf{I}_k$ and $\boldsymbol{\alpha}_k = \beta_{k,j} = 0$ for any index k and j.

Proof: From (9–10), when $\mathbf{A}_k = \mathbf{I}_k$ and $\boldsymbol{\alpha}_k = \beta_{k,j} = 0$, we should minimize the following quantity:

$$\begin{aligned} &||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{A}_k \mathbf{U}_k^T \hat{\mathbf{X}}_k||^2 + \boldsymbol{\alpha}_k ||\mathbf{U}_k||^2 + \sum_j \beta_{kj} |\mathbf{u}_{kj}| \\ &= ||\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k - \mathbf{I}_k \mathbf{U}_k^T \hat{\mathbf{X}}_k||^2 \\ &= tr(\mathbf{\Phi}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{\Phi}_k + \mathbf{U}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k - 2\mathbf{U}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{\Phi}_k) \end{aligned}$$

This minimization problem converts to

$$\max tr(\mathbf{U}_k^T \hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T \mathbf{U}_k)$$
$$\textit{subject to } \mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_k$$

Since when $\mu = -1$, $\hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T = \mathbf{S}_T^{(k)}$. This gives the conclusion. ■

Theorem 5 shows that MPCA is a special case of STDA. In facts, the MPCA proposed in [25] and CSA in [24] are exactly the same in essence except for the notations and statements. Therefore, according to theorem 5 and [24], it is easy to have the following conclusions:

Corollary 3: If $n = 1$ (i.e. for vector input), PCA is equivalent to the special case of STDA when $\mu = -1$, $\mathbf{A}_1 = \mathbf{U}_1$ and $\boldsymbol{\alpha}_1 = \beta_{1,j} = 0$ for any index j.

Corollary 4: If $n = 1$ (i.e. for vector input), SPCA is equivalent to the special case of STDA when $\mu = -1$.

Corollary 5: If $n = 2$ (i.e. for image input), 2DPCA is equivalent to the special case of STDA when $\mu = 1$, $\mathbf{A}_1 = \mathbf{U}_1$, $\mathbf{A}_2 = \mathbf{U}_2 = I_2$ and $\boldsymbol{\alpha}_k = \beta_{k,j} = 0$ for any index k and j.

B. Connection to MMC-Based Methods

MMC was extended to the multilinear case, named TMMC in [31]. In fact, TMMC is a special case of GTDA proposed in [29], i.e. if $\mu = 1$, then GTDA reduces to TMMC. The objective function of GTDA is as follows:

$$\mathbf{U}_k^* = \arg\max tr(\mathbf{U}_k^T (\mathbf{S}_B^{(k)} - \mu \mathbf{S}_W^{(k)}) \mathbf{U}_k)$$
$$\textit{subject to } \mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_k$$

The optimal projection of the mode-k unfolding of the tensors can be obtained by performing SVD or solving the standard eigen-function $(\mathbf{S}_B^{(k)} - \mu \mathbf{S}_W^{(k)})\mathbf{U}_k = \mathbf{\Lambda}_k \mathbf{U}_k$. The following theorem shows the close relationship between GTDA and STDA when the order of the projections is neglected.

Theorem 6: The projections of TMMC are the same as the projections of STDA when $\mathbf{A}_k = \mathbf{I}_k$ and $\boldsymbol{\alpha}_k = \beta_{k,j} = 0$ for any index k and j.

Proof: The proof is similar to the proof of Theorem 5 except for letting $\hat{\mathbf{X}}_k \hat{\mathbf{X}}_k^T = \mathbf{S}_B^{(k)} - \mu \mathbf{S}_W^{(k)}$. ■

Theorem 6 can result in the following conclusions.

Corollary 6: If $n = 1$ (i.e. for vector input), MMC is equivalent to the special case of STDA when $\mathbf{A}_k = \mathbf{I}_k$ and $\boldsymbol{\alpha}_1 = \beta_{1,j} = 0$ for any index j.

Corollary 7: If $n = 2$ (i.e. for image input), the bilateral 2DMMC [49] is equivalent to the special case of STDA when $\boldsymbol{\alpha}_k = \beta_{k,j} = 0$ for any index k and j.

IV. Experiments

In this section, a set of experiments are presented to evaluate the proposed STDA algorithm for face image recognition tasks using second order tensor and action recognition using high (third) order tensor. The Yale, FERET CMU PIE face databases were used to test the performance or robustness of STDA with variations in face expression, pose and lighting conditions. The Weizmann database was used to test the performance of STDA in high-order tensor learning. The nearest neighbor classifier with Euclidean distance was used in all the experiments.

Fig. 1. The sample images of one person from the Yale face database.

A. Exploration on Yale Database

The Yale face database (http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 images of 15 individuals (each person providing 11 different images) with various facial expressions and lighting conditions. In our experiments, each image was manually cropped and resized to 50×40 pixels. Fig. 1 shows sample images of one person on the Yale face database.

Explore the performance of the parameters: in order to explore the variations of the performance of the STDA against the parameters, 4 images per individual were randomly selected for training and the remaining images were used as test in the experiment. The variation of the recognition rate versus the sparseness parameter cardinality K in STDA on Yale face database is shown in Fig. 2 (a), from which it can be found that when $K \in [9, 10, 11]$ STDA achieves its best performance. Fig. 2 (b) and (c) shows the variation of the recognition rate versus the values of $\boldsymbol{\alpha}$ (Alpha) and μ (Miu), respectively. In the experiments, the recognition rates of STDA is robust to parameter $\boldsymbol{\alpha}$ in a larger range from 10^{-4} to 10^3. Fig. 2 (c) shows that STDA achieves its best performance when $\mu = 10^3$. Similar performances can also be found in other databases used in this paper.

Experimental setting: In the experiments, 4 images of each individual were randomly selected and used as training set, and one half of the remaining images as validation set and test set, respectively. The experiments were independently performed 10 times and the average recognition results on the test set were calculated. For each run, the validation set was used for parameters' selection. The optimal sparseness parameter cardinality K and the optimal subspace dimensions were ranged in [1], [40]. When Elastic Net was used, the $\boldsymbol{\alpha}_k$'s were selected from $10^{-4}, 10^{-3}, \ldots, 10^5$, and the parameters $\beta_{k,j}$s can be automatically determined since the Elastic Net algorithm could provide the optimal solution path of $\beta_{k,j}$s for given $\boldsymbol{\alpha}_k$ [38]. And the coefficients of the L_1 norm in SNTF [45] and HNTF [44], and μ in STDA were all selected from $10^{-4}, 10^{-3}, \ldots, 10^5$. For SLDA [42], both the L_1 and L_2 norms penalty weights were also selected from $10^{-4}, 10^{-3}, \ldots, 10^5$. For MPCA, MLDA, TMMC, the main parameters are the dimensions of each subspaces. For each subspace, the dimensions were varied from 1 to the size of the image matrix with step equaling to 1. Thus all the parameter ranges/combinations are covered. When using the strategy of "multilinear methods plus LDA", PCA was used for preprocessing after multilinear feature extraction so that the within-class scatter matrix is invertible (LDA is more stable in this case). For each run, the optimal parameters determined by the validation set were used in the algorithms.

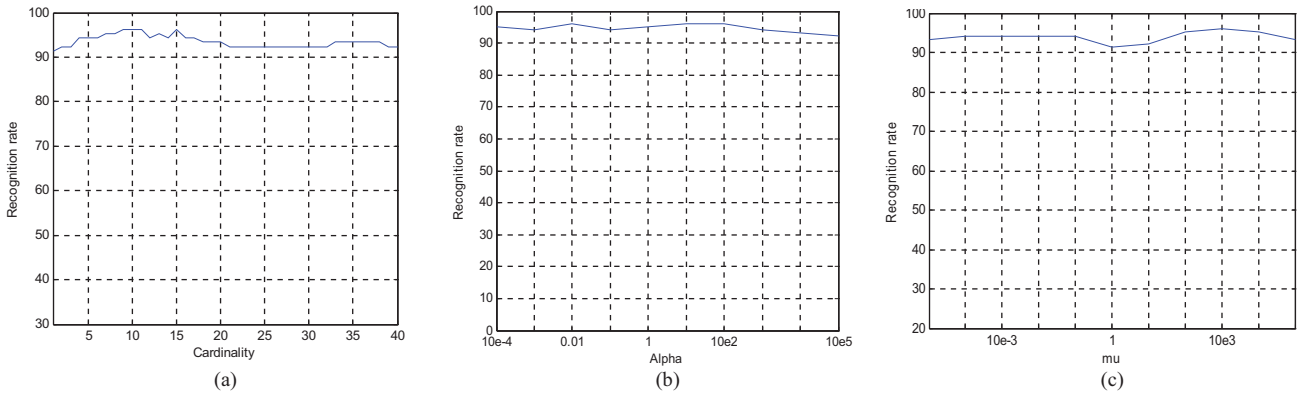

Fig. 2. (a) The recognition rate (%) vs. the cardinality. (b) The recognition rate vs. the value of Alpha. (c) The average recognition rates vs. the value of Miu.

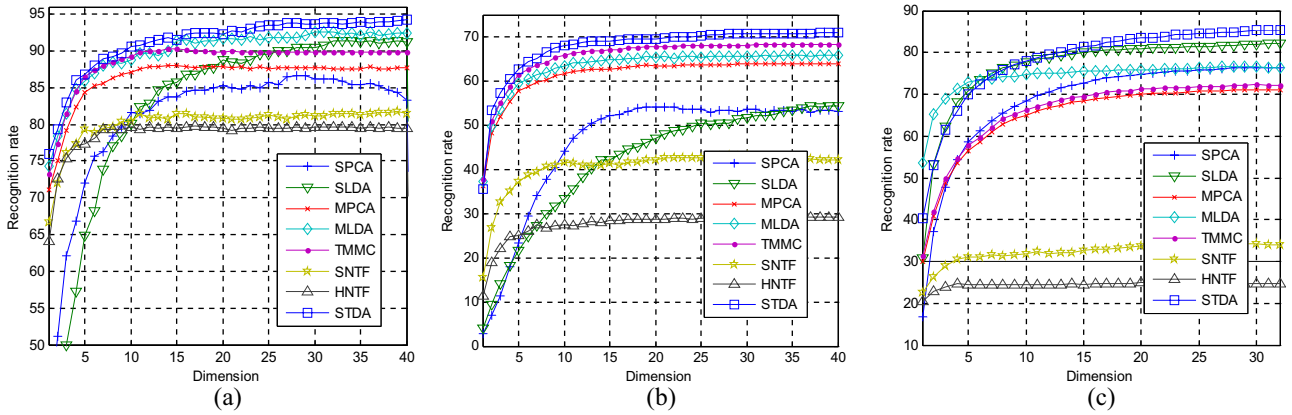

Fig. 3. The average recognition rates (%) versus the dimensions on the Yale (a), FERET (b) and CMU PIE (c) face databases.

The average recognition rates of the test set and the corresponding dimensions and standard deviations of each method are shown in Table II. The recognition rates vs. the variations of the dimension are shown in Fig. 3 (a), in which the dimension (i.e. the horizontal axis) means the number of the column vectors in the projection matrices used for feature extraction. As can be seen form Table II and Fig. 3 (a), STDA obtains the best recognition rates in the experiments, which shows the robustness for the variations on facial expressions and lighting conditions.

B. Experiments on FERET Face Database

The FERET face database is a result of the FERET program, which was sponsored by the US Department of Defense through the DARPA Program [44]. It has become a standard database for testing and evaluating state-of-the-art face recognition algorithms. The proposed method was tested on a subset of the FERET database. This subset includes 1,400 images of 200 individuals (each individual has seven images) and involves variations in facial expression, illumination, and pose. In the experiment, the facial portion of each original image was automatically cropped based on the location of the eyes, and the cropped images was resized to 40×40 pixels. The sample images of one person are shown in Fig. 4.

In the experiments, $l(l = 3, 4, 5)$ images of each individual were randomly selected and used for training, and one half of the rest images were used for validation and test, respectively. The experiments were performed as the same way in Section A. Table III lists the recognition rates of each method and Fig. 3 (b) shows the recognition rates vs. the variations of the dimensions. Again, STDA performs better than the other methods.

Fig. 4. Sample images of one person on FERET face database.

C. Experiments on CMU PIE Face Database

The CMU PIE face database [50] contains 68 individual with 41,368 face images as a whole. The face images were captured under varying pose, illumination and expression. In our experiments, we select a subset (C29) which contains1632 images of 68 individuals (each individual has 24 images).The C29 subset involves variations in illumination, facial expression and pose. All of these face images are aligned based on eye coordinates and cropped to 32×32. Fig. 5 shows the sample images from this database.

In the experiments, $l(l = 3, 4, 5)$ images of each individual were randomly selected and used as training set,

TABLE II
THE PERFORMANCE OF DIFFERENT METHODS ON YALE FACE DATABASE

Method	SPCA	SLDA	MPCA	MPCA +LDA	MLDA	MLDA +LDA	TMMC	TMMC +LDA	SNTF	SNTF +LDA	HNTF	HNTF +LDA	STDA	STDA +LDA
Recognition rate	86.63	87.38	88.07	92.75	89.56	94.19	90.24	94.33	81.77	89.22	79.70	89.81	93.30	95.59
Standard deviation	±3.27	±2.52	±2.79	±3.30	±3.02	±2.91	±3.12	±2.68	±5.32	±3.23	±4.84	±2.19	±2.47	±2.13
Dimension	28	39	14×17	14	15×28	14	14×16	14	44×39	14	17×16	14	43×21	14

TABLE III
THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF DIFFERENT METHODS ON FERET FACE DATABASE

Training Samples	SPCA	SLDA	MPCA	MPCA +LDA	MLDA	MLDA +LDA	TMMC	TMMC +LDA	SNTF	SNTF +LDA	HNTF	HNTF +LDA	STDA	STDA +LDA
3	46.43	46.17	51.17	57.84	52.87	61.41	54.29	64.06	34.84	55.41	23.98	52.25	57.86	69.54
	115	195	35×16	199	38×19	199	35×16	199	27×23	199	36×30	199	36×16	199
	±11.55	±10.18	±9.00	±10.50	±7.38	±6.83	±10.17	±9.67	±6.53	±9.75	±3.57	±8.75	±5.68	±6.60
4	54.16	54.33	63.85	71.71	65.93	73.66	68.24	73.52	43.05	69.62	29.08H	65.38	70.95	75.68
	95	185	32×16	199	36×32	199	32×16	199	28×30	199	30×30	199	38×38	199
	±10.52	±6.83	±7.09	±8.70	±7.02	±7.90	±8.09	±7.62	±7.44	±11.09	±5.89	±10.81	±5.11	±5.44
5	58.65	60.44	72.23	79.37	74.52	81.45	75.45	82.14	44.15	76.20	29.18H	72.12	78.56	85.77
	90	199	37×19	199	36×22	199	37×16	199	39×30	199	35×35	199	32×16	199
	±7.05	±6.12	±8.06	±6.04	±9.05	±5.96	±7.07	±5.09	±13.03	±9.55	±8.74	±7.14	±5.03	±4.08

Fig. 5. The sample images of one person form the CMU PIE face database.

and one half of the remaining images as validation and test set, respectively. The experimental parameters were set as in Section A. The performances of each method are shown in Table IV. The recognition rates vs. the variations of the dimension (5 samples were used for training) are shown in Fig. 3 (c). As can be seen form Table IV and Fig. 3 (c), STDA obtains the best recognition rates in all the cases when there are variations in expression, pose and illumination.

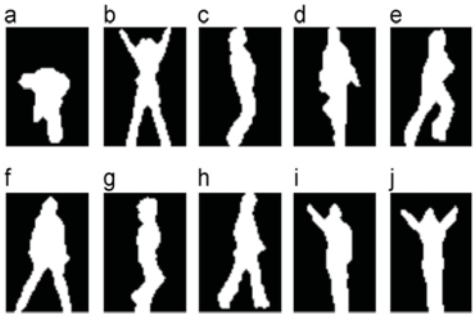

Fig. 6. Key silhouettes of 10 actions from the Weizmann database. (a) bend, (b) jack, (c) jump, (d) pjump, (e) run, (f) side, (g) skip, (h) walk, (i) wave1, (j) wave2.

D. Experiments on Weizmann Action Database

The experiment was performed on the Weizmann database [45], which was a commonly used database for human action recognition. The 90 videos coming from 10 categories of actions included bending (bend), jacking (jack), jumping (jump), jumping in places (pjump), running (run), galloping-side ways (side), skipping (skip), walking (walk), single-hand waving (wave1), and both-hands waving (wave2), which were performed by nine subjects. The centered key silhouettes of each action are shown in Fig. 6.

In order to represent the spatiotemporal feature of the samples, 10 successive frames of each action were used to extract the temporal feature. Fig. 7 shows a tensor sample of the bending action. Each centered frame was normalized to the size of 32×24 pixels. Thus the tensor sample was represented in the size of $32 \times 24 \times 10$ pixels. It should be note that there is no overlapped frames in any two tensors and the starting frames of the tensors are not normalized to the beginning frames of each action. Thus, the recognition tasks are difficult and close to the real-world applications. Therefore, if one wants to get high recognition accuracy, the methods used for feature extraction should be robust to starting frames and actions' variations.

In the experiments, 6 action tensors of each category were randomly selected and used for training and one half of the remaining tensors as validation and test set, respectively. The experimental procedures were the same as in Section A. The recognition rates of each method are listed in Table V, and the variations of the average recognition rates versus the dimensions are shown in Fig. 8. It can be found that STDA also outperforms the other algorithms in action tensor feature extraction.

TABLE IV

THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF 5 METHODS ON THE CMU PIE FACE DATABASE

Training Samples	SPCA	SLDA	MPCA	MPCA +LDA	MLDA	MLDA +LDA	TMMC	TMMC +LDA	SNTF	SNTF +LDA	HNTF	HNTF +LDA	STDA	STDA +LDA
3	53.66	62.56	48.02	64.79	57.32	64.26	49.36	67.69	23.72	63.94	19.24	65.65	63.90	70.18
	140	145	29×29	67	28×28	67	31×27	67	25×6	67	30×26	67	31×21	67
	±10.93	±10.67	±5.41	±10.23	±11.05	±10.49	±5.42	±7.92	±5.52	±11.81	±3.03	±10.91	±7.18	±6.84
4	68.00	75.49	61.30	75.43	68.24	75.09	62.64	78.86	27.99	78.09	21.55	77.57	76.55	82.21
	145	150	29×30	67	24×21	67	30×27	67	16×6	67	29×21	67	31×28	67
	±7.87	±6.71	±9.45	±14.33	±3.02	±15.25	±7.47	±7.79	±5.86	±8.81	±4.04	±8.96	±5.66	±5.09
5	76.39	81.61	71.08	83.12	76.67	82.94	72.26	85.89	34.25	85.75	24.87	84.70	85.41	88.09
	150	155	31×26	67	28×28	67	29×29	67	30×26	67	23×21	67	31×28	67
	±7.76	±3.66	±8.87	±3.31	±3.25	±6.77	±3.07	±7.14	±7.47	±8.49	±4.45	±6.04	±4.17	±3.46

TABLE V

THE PERFORMANCE OF DIFFERENT METHODS ON THE WEIZMANN ACTION DATABASE

Method	SPCA	SLDA	MPCA	MPCA +LDA	MLDA	MLDA +LDA	TMMC	TMMC +LDA	SNTF	SNTF +LDA	HNTF	HNTF +LDA	STDA	STDA +LDA
Recognition rate	78.03	76.59	70.14	76.24	76.94	75.87	77.19	76.12	69.21	61.47	48.93	56.60	80.38	80.77
Standard deviation	±3.20	±2.55	±2.83	±2.08	±3.68	±3.94	±3.54	±2.08	±4.06	±2.62	±2.64	±3.53	±2.98	±2.39
Dimension	16	40	10^3	9	9^3	9	9^3	9	9^3	9	9^3	9	10^3	9

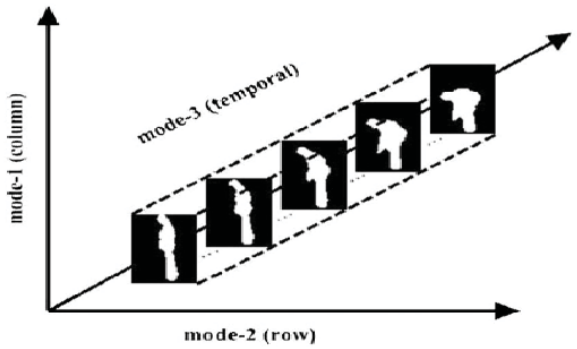

Fig. 7. An example of the bending action in spatiotemporal domain from Weizmann database.

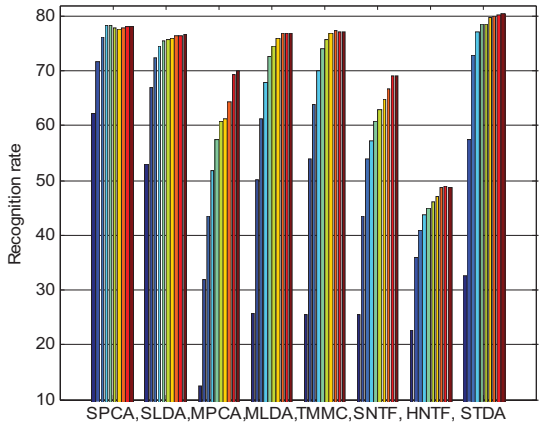

Fig. 8. The average recognition rates (%) versus the dimensions on the Weizmann action database.

E. Observations and Discussions

Based on the experimental results shown in above sections, the following observations are obtained:

(1) It can be found from the experiments that the sparse learning algorithm such as SPCA and SLDA can even perform better than MPCA and MLDA, respectively. Although the label information was not used in SPCA, SPCA performs better than MLDA and obtains high recognition rates in action recognition. STDA performed better than SPCAn and SLDAn, which indicates that combining the L_1 and L_2 norms for sparse tensor learning with higher order data can obtain more discriminative information than the simple linear cases.

(2) STDA uses the same criterion (i.e. the differential form) as in TMMC, but the recognition rates of STDA are significantly higher than the ones obtained by TMMC. This indicates that introducing the sparsity in the projection vectors/matrices can enhance the performance of the criterion used in the algorithm. The reason is that using the L_1 norm for sparse discriminant learning can avoid the over-fitting and thus it can be more robust or obtain more generalization abilities than the related methods.

(3) For the tensor learning methods, MLDA and TMMC usually performs better than MPCA. With the similar criterion by adding the L_1 and L_2 norms terms for sparse learning in STDA, the performance was also enhanced. Similar cases also happen on different databases. This proves that introducing the sparsity constraint to learn the discriminant subspace is a tractable method for improving the performance when there are variations in expression, pose and illumination in the face images and the starting point in the action tensor.

(4) Since SNTF and HNTF aim to find the nonnegative semantic structure information embedding in the data

instead of feature extraction for classification, they obtain low recognition rates. But it can be greatly enhanced by using LDA for further dimensionality reduction. Usually, the strategy of "multilinear methods plus LDA" can significantly enhance the performance.

V. Conclusion

A sparse tensor learning method called STDA was designed in this paper for feature extraction. The L_1 and L_2 norms were integrated to the maximal marginal criterion and thus a novel sparse learning model was proposed. The optimal solutions of this model can be obtained by the iterative algorithm using the Elastic Net regression. Theoretical analyses were presented to explore the properties of STDA. The relationships between STDA and other algorithms were also shown. Experimental results showed that STDA performs better than the well-known sparse linear dimensionality reduction algorithms and the extensions of the classical multilinear subspace learning methods in face recognition and action recognition. It is shown that STDA is more robust than the compared methods on the variations in expression, pose and illumination in the face images and the different starting frames and variations in the action tensor. Usually, STDA with higher order data can obtain more discriminative information than the simple linear cases.

Appendix

A. Proof of Theorem 3

Proof:

$$||\mathbf{\Phi}_k^T\hat{\mathbf{X}}_k - \mathbf{A}_k\mathbf{U}_k^T\hat{\mathbf{X}}_k||^2 + \boldsymbol{\alpha}_k||\mathbf{U}_k||^2 = tr(\mathbf{\Phi}_k^T\hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^T\mathbf{\Phi}_k - 2\mathbf{U}_k^T\hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^T\mathbf{\Phi}_k\mathbf{A}_k + \mathbf{U}_k^T\hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^T\mathbf{U}_k + \boldsymbol{\alpha}\mathbf{U}_k^T\mathbf{U}_k). \quad \text{(A1)}$$

For the fixed $\mathbf{A}_k$, using Lagrange multiplier method, the above quantity is minimized at

$$\hat{\mathbf{U}}_k = (\hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^T + \boldsymbol{\alpha}\mathbf{I}_k)^{-1}\hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^T\mathbf{\Phi}_k\mathbf{A}_k = \mathbf{\Phi}_k\mathbf{\Lambda}_k(\mathbf{\Lambda}_k + \boldsymbol{\alpha}\mathbf{I}_k)^{-1}\mathbf{A}_k \quad \text{(A2)}$$

On the other hand, for the given $\mathbf{U}_k$, minimizing (A1) gives

$$\arg\min tr(-\mathbf{U}_k^T\hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^T\Phi_k\mathbf{A}_k) \quad s.t. \quad \mathbf{A}_k^T\mathbf{A}_k = \mathbf{I}_k \quad \text{(A3)}$$

Then minimizing (A3) is equivalent to the following maximizing problem:

$$\arg\max A_k^T\mathbf{\Lambda}_k^2(\mathbf{\Lambda}_k + \boldsymbol{\alpha}\mathbf{I}_k)^{-1}\mathbf{A}_k, s.t. \quad A_k^T\mathbf{A}_k = \mathbf{I}_k \quad \text{(A4)}$$

Since this is a standard eigen-decomposition problem, the solution can be obtained by SVD of $\mathbf{\Lambda}_k^2(\mathbf{\Lambda}_k + \boldsymbol{\alpha}\mathbf{I}_k)^{-1}$:

$$\mathbf{\Lambda}_k^2(\mathbf{\Lambda}_k + \boldsymbol{\alpha}\mathbf{I}_k)^{-1} = \bar{\mathbf{U}}_k\bar{D}_k\bar{V}_k^T \quad \text{(A5)}$$

Then $\hat{A}_k = \bar{\mathbf{U}}_k$ is the optimal solution of (A4). Therefore, it can be seen from (A2) that $\hat{\mathbf{U}}_k$ always spans the same subspace in iterative procedures as the one spanned by $\mathbf{\Phi}_k$. Thus $\mathbf{u}_k^j \propto \varphi_{kj}(1 \le j \le d_k)$. ■

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriengman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *J. Amer. Stat. Assoc.*, vol. 97, no. 1, pp. 611–631, 2002.

[3] Q. Liu, H. Lu, and S. Ma, "Improving kernel Fisher discriminant analysis for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 42–49, Jun. 2004.

[4] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature suvey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[5] L. Sirovich and M. Kirby, "Low-dimensional procedure for characterization of human faces," *J. Opt. Soc. Amer.*, vol. 4, no. 3, pp. 519–524, Mar. 1987.

[6] M. Kirby and L. Sirovich, "Application of the Karhunen–Loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[7] M. Turk, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 10, no. 9, pp. 358–386, Jan. 1991.

[8] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[9] J. Yang, D. Zhang, X. Yong, and J. Yang, "Two-dimensional discriminant transform for face recognition," *Pattern Recognit.*, vol. 38, no. 7, pp. 1125–1129, Jul. 2005.

[10] M. Li and B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recognit. Lett.*, vol. 26, no. 5, pp. 527–532, Apr. 2005.

[11] J. Ye, "Generalized low rank approximations of matrices," *Mach. Learn.*, vol. 61, nos. 1–3, pp. 167–191, Nov. 2005.

[12] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," *Adv. Neural Inf. Process. Syst.*, vol. 17, pp. 1569–1576, Jul. 2004.

[13] W. Zuo, D. Zhang, and K. Wang, "Bidirectional PCA with assembled matrix distance metric for image recognition.," *IEEE Trans. Syst., Man, Cybern. Part B, Cybern.*, vol. 36, no. 4, pp. 863–872, Aug. 2006.

[14] W. Zuo, D. Zhang, J. Yang, and K. Wang, "BDPCA plus LDA: A novel fast feature extraction technique for face recognition," *IEEE Trans. Syst., Man, Cybern. Part B, Cybern.*, vol. 36, no. 4, pp. 946–952, Apr. 2006.

[15] D. Zhang and Z. Zhou, "$(2D)^2$PCA: Two-directional Two-dimensional PCA for efficient face representation and recogntion," *Neurocomputing*, vol. 69, nos. 1–3, pp. 224–231, Dec. 2005.

[16] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, and H. Zhang, "Human gait recognition with matrix representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 896–903, Jul. 2006.

[17] T. G. Kolda, "Orthogonal tensor decompositions," *SIAM J. Matrix Anal. Appl.*, vol. 23, no. 1, pp. 243–255, 2001.

[18] L. Lathauwer, B. D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[19] L. Lathauwer, B. D. Moor, and J. Vandewalle, "On the best rank-1 and rank-(R1,R2,...RN) approximation of high-order tensors," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1324–1342, 2000.

[20] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensiofaces," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 447–460.

[21] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear image analysis for facial recognition," in *Proc. 16th Int. Conf. Pattern Recognit.*, vol. 2. 2002, pp. 511–514.

[22] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear subspace analysis for image ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. 93–99.

[23] S. Rana, W. Liu, M. Lazarescu, and S. Venkatesh, "A unified tensor framework for face recognition," *Pattern Recognit.*, vol. 42, no. 11, pp. 2850–2862, Nov. 2009.

[24] D. Xu, S. Yan, L. Zhang, S. Lin, H. Zhang, and T. S. Huang, "Reconstruction and recogntition of tensor-based objects with concurrent subspaces analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 36–47, Jan. 2008.

[25] H. Lu, K. N. K. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 18–39, Jan. 2008.

[26] H. Lu, K. N. K. Plataniotis, and A. N. Venetsanopoulos, "Uncorrelated multilinear principal component analysis for unsupervised multilinear subspace learning," *IEEE Trans. Neura Netw.*, vol. 20, no. 11, pp. 1820–1836, Nov. 2009.

[27] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 212–220, Jan. 2007.

[28] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Uncorrelated multilinear discriminant analysis with regularization and aggregation for tensor object recognition," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 103–123, Jan. 2009.

[29] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–17015, Oct. 2007.

[30] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York, NY, USA: Academic, 1990.

[31] R.-X. Hu, W. Jia, D.-S. Huang, and Y.-K. Lei, "Maximum margin criterion with tensor representation," *Neurocomputing*, vol. 73, nos. 10–12, pp. 1541–1549, Jun. 2010.

[32] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006.

[33] Y. Wang and S. Gong, "Tensor discriminant analysis for view-based object recognition," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3. 2006, pp. 33–36.

[34] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "A survey of multilinear subspace learning for tensor data," *Pattern Recognit.*, vol. 44, no. 7, pp. 1540–1551, Jul. 2011.

[35] S. Lin and T. S. Huang, "Reconstruction and recognition of tensor-based objects with concurrent subspaces analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 36–47, Jan. 2008.

[36] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[37] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soci., Ser. B (Stat. Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[38] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal Stat. Soc., Ser. B (Stat. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[39] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.

[40] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, Jun. 2006.

[41] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[42] Z. Qiao, L. Zhou, and J. Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data," *IAENG Int. J. Appl. Math.*, vol. 39, no. 1, pp. 48–60, Jan. 2009.

[43] R. Sivalingam, D. Boley, and V. Morellas, "Tensor sparse coding for region covariances," in *Proc. Comput. Vis. ECCV*, vol. 4. 2010, pp. 722–735.

[44] M. Mørup, L. K. Hansen, and S. M. Arnfred, "Algorithms for sparse nonnegative Tucker decompositions," *Neural Comput.*, vol. 20, no. 8, pp. 2112–2131, Aug. 2008.

[45] J. Liu, J. Liu, P. Wonka, and J. Ye, "Sparse non-negative tensor factorization using columnwise coordinate descent," *Pattern Recognit.*, vol. 45, no. 1, pp. 649–656, Jan. 2012.

[46] X. Li, S. Member, S. Lin, S. Yan, and D. Xu, "Discriminant locally linear embedding with high-order tensor data," *IEEE TTrans. Syst., Man, Cybern. Part B, Cybern.*, vol. 38, no. 2, pp. 342–352, Apr. 2008.

[47] Y. Liu, Y. Liu, and K. C. C. Chan, "Tensor distance based multilinear locality-preserved maximum information embedding," *IEEE Trans. Neural Netw.*, vol. 21, no. 11, pp. 1848–1854, Nov. 2010.

[48] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang, "Marginal Fisher analysis and its variants for human gait recognition and content-based image retrieval," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2811–2821, Nov. 2007.

[49] Q. Gu and J. Zhou, "Two dimensional maximum margin criterion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 1621–1624.

[50] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

Zhihui Lai received the B.S. degree in mathematics from South China Normal University, Guangzhou, China, the M.S. degree from Jinan University, Guangzhou, China, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2002, 2007, and 2011, respectively. He has been a Research Associate with the Hong Kong Polytechnic University, Hong Kong, since 2010. Currently, he is a Post-Doctoral Fellow with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Harbin, China. He is the author of more than 30 scientific papers in pattern recognition and computer vision. His current research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization, and applications in the fields of intelligent robot research.

Yong Xu (M'06) received the B.S. and M.S. degrees from the Air Force Institute of Meteorology, Nanjing, China, in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, in 2005. From May 2005 to April 2007, he was with the Shenzhen Graduate School, Harbin Institute of Technology (HIT), Harbin, China, as a Post-Doctoral Research Fellow. He is currently an Associate Professor with the Shenzhen Graduate School, HIT. He was a Research Assistant with the Hong Kong Polytechnic University, Hong Kong, from August 2007 to June 2008. His current interests include pattern recognition, biometrics, and machine learning. He has published more than 40 scientific papers.

Jian Yang received the B.S. degree in mathematics from Xuzhou Normal University, Xuzhou, China, in 1995, the M.S. degree in applied mathematics from Changsha Railway University, Changsha, China, in 1998, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2002. In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza, Zaragoza, Spain, and received the RyC Program Research Fellowship sponsored by the Spanish Ministry of Science and Technology. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre of the Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Professor with the School of Computer Science and Technology, NUST. He is the author of more than 50 scientific papers in pattern recognition and computer vision. His current research interests include pattern recognition, computer vision, and machine learning.

Jinhui Tang is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. He received the B.E. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. From 2008 to 2010, he was a Research Fellow with the School of Computing, National University of Singapore, Singapore. He was with the School of Information and Computer Science, UC Irvine, Irvine, CA, USA, from 2010 to 2010, as a Visiting Research Scientist. From 2011 to 2012, he was with Microsoft Research Asia, Beijing, China, as a Visiting Researcher. His current research interests include large-scale multimedia search, social media mining, and computer vision. He has authored over 80 journal and conference papers. He serves as an Editorial Board Member of *Pattern Analysis and Applications*, *Multimedia Tools and Applications*, *Information Sciences*, and *Neurocomputing*, a Technical Committee Member at over 30 international conferences, and a Reviewer for over 30 prestigious international journals. He was a co-recipient of the Best Paper Award form ACM Multimedia in 2007, PCM in 2011, and ICIMCS in 2011. He is a member of ACM and CCF.

David Zhang (F'08) received the Degree in computer science from Peking University, Beijing, China, and the M.Sc. degree in computer science and the Ph.D. degree from the Harbin Institute of Technology (HIT), Harbin, China, in 1982 and 1985, respectively. From 1986 to 1988, he was a Post-Doctoral Fellow with Tsinghua University Beijing, and an Associate Professor with Academia Sinica, Beijing. In 1994, he received the second Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada. Currently, he is the Head of the Department of Computing and a Chair Professor with the Hong Kong Polytechnic University, Hong Kong. He serves as a Visiting Chair Professor with Tsinghua University and an Adjunct Professor with Peking University, Shanghai Jiao Tong University, Shanghai, China, HIT, and the University of Waterloo. He is the Founder and Editor-in-Chief of *International Journal of Image and Graphics*, Book Editor, Springer International Series on Biometrics, Organizer, the International Conference on Biometrics Authentication, an Associate Editor of more than ten international journals, including the IEEE TRANSACTIONS AND PATTERN RECOGNITION, and the author of more than ten books and 200 journal papers. He is a Croucher Senior Research Fellow, Distinguished Speaker of the IEEE Computer Society, and a Fellow of both IEEE and IAPR.