Joint Tensor Feature Analysis For Visual Object Recognition

Wai Keung Wong, Zhihui Lai, Yong Xu, *Member, IEEE,* Jiajun Wen, and Chu Po Ho

***Abstract*—Tensor-based object recognition has been widely studied in the past several years. This paper focuses on the issue of joint feature selection from the tensor data and proposes a novel method called joint tensor feature analysis (JTFA) for tensor feature extraction and recognition. In order to obtain a set of jointly sparse projections for tensor feature extraction, we define the modified within-class tensor scatter value and the modified between-class tensor scatter value for regression. The k-mode optimization technique and the $L_{2,1}$-norm jointly sparse regression are combined together to compute the optimal solutions. The convergent analysis, computational complexity analysis and the essence of the proposed method/model are also presented. It is interesting to show that the proposed method is very similar to singular value decomposition on the scatter matrix but with sparsity constraint on the right singular value matrix or eigen-decomposition on the scatter matrix with sparse manner. Experimental results on some tensor datasets indicate that JTFA outperforms some well-known tensor feature extraction and selection algorithms.**

***Index Terms*—Discriminant analysis, feature selection, object recognition, sparse learning.**

I. Introduction

FEATURE extraction or feature selection is an important issue in pattern recognition. The classical feature extraction methods such as principle component analysis (PCA) [1], [2] and linear discriminant analysis (LDA) [3]–[5] are frequently used in the fields of pattern recognition, computer vision, and data mining. PCA focuses on minimizing the error of the reconstructive information while LDA is used to maximize the discriminative information with respect to the classification.

In past decades, there were many extended versions of PCA and LDA. One of the key developments is the tensor extensions of these classical methods. For example, two dimensional PCA (2-DPCA) proposed by Yang *et al.* [6], bidirectional PCA proposed by Zuo *et al.* [7], image-based generalized low rank approximation proposed by Ye [8] and the multilinear PCA by Lu *et al.* [9] are the representative methods on the developments of the classical PCA. Similarly, LDA was also extended to be two dimensional LDA (2-DLDA) [10] and multilinear LDA (MLDA) [11] for image feature extraction and color face recognition [12], [13].

In order to avoid the singularity in computing the inverse matrix of the within-class scatter matrix, maximal margin criterion (MMC) [14] and its variation [15] were proposed for discriminant feature extraction. MMC was also extended to second order case [16], [17] and high order tensor cases, i.e., general tensor discriminant analysis (GTDA) [18] and tensor MMC (TMMC) [19], for feature extraction on the visual object.

Some other interesting applications based on the tensor representations were also explored. For example, the empirical discriminative tensor analysis for crime forecasting [20], which minimizes the empirical risk and preserves the discriminative information. The biologically inspired features [21], [22] based on the tensor data can also be used for face and gait recognition. Tao *et al.* [23] proposed the supervised tensor learning (STL) framework which is the multilinear extension of the convex optimization-based learning algorithm. Multivariate multilinear regression (MMR) [24] was applied to model the fitting procedure in the active appearance model [25]. For the comprehensive understanding of the tensor learning methods, readers can refer to the survey [26] for more details.

However, these classical methods and its high order extensions cannot obtain the sparse projection for feature extraction and selection. Recently, sparse subspace learning methods has been paid much attention on feature extraction. The common property of these methods is to use the L_1 norm-based sparse regression methods [27]–[29] so as to learn the sparse projections. By using these techniques, the classical methods,

Manuscript received April 11, 2014; revised August 13, 2014 and October 31, 2014; accepted November 11, 2014. Date of publication December 9, 2014; date of current version October 13, 2015. This work was supported in part by the Natural Science Foundation of China under Grant 61203376, Grant 61375012, Grant 61203247, Grant 61071179, Grant 61362031, Grant 61332011, and Grant 61370163, in part by the Hong Kong Polytechnic University Project under Grant G-YJ22, in part by the China Post-Doctoral Science Foundation under Project 2012M510958 and Project 2013T60370, in part by the Guangdong Natural Science Foundation under Project S2012040007289, and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant JC201005260122A, Grant JCYJ201206 13153352732, Grant JCYJ20120613134843060, and Grant JCYJ20130329152024199. This paper was recommended by Associate Editor D. Tao. *(Corresponding author: Z. Lai.)*

W. K. Wong is with the Institute of Textiles and Clothing, Hong Kong Polytechnic University, Hong Kong, and also with Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518055, China.

Z. Lai is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China, also with the Institute of Textiles and Clothing, Hong Kong Polytechnic University, Hong Kong, and also with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: lai_zhi_hui@163.com).

Y. Xu and J. Weng are with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China.

C. P. Ho is with the Institute of Textiles and Clothing, Hong Kong Polytechnic University, Hong Kong.

The first two authors contributed equally to this work.

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/TCYB.2014.2374452

i.e., PCA and LDA, were extended to sparse PCA (SPCA) [30] and sparse discriminant analysis (SDA) [31]. It is shown that by introducing the L_1 norm-based sparseness constraint on the learned subspace, the function of simultaneous feature selection and dimensionality reduction can be achieved by these sparse feature extraction methods, which lead to good performance in the recognition task and are more robust to the outliers. Therefore, many types of SPCA [32], [33] and the modified versions of sparse LDA [34]–[36] were proposed in recent years.

No matter the L_1 norm-based constraint or the directly cardinality constraint on the bases of the learned subspace, a potential drawback (i.e., SPCA, SDA, and their extended versions mentioned above) is that they cannot provide the subspaces with joint (consistent) sparseness. To address this problem, joint $L_{2,1}$-norms regularization method was proposed in [37] for robust feature selection. By using the same technique, Gu *et al.* [38] proposed the feature selection and subspace learning (FSSL) method combining the graph spectral analysis and joint $L_{2,1}$-norm regularization. Similarly, Hou *et al.* [39] integrated the $L_{2,1}$-norm regression and spectral decomposition to obtain the jointly sparse subspace for feature extraction. Since the $L_{2,1}$-norm-based regularized feature selection can jointly select the most relevant features from the data points and is more robust than the traditional L_2 norm regularized techniques, joint feature selection has been used in semi-supervised leaning for multimedia data understanding [40], automatic image annotation [41], and classifier design [42].

Although the $L_{2,1}$-norm-based methods have been widely used in many fields, most of the previous works focus on the high dimensional vector-based representation. Due to the potential singularity in computing the inverse of the matrix and the heavy computational burdens caused by the very high dimensionality patterns in optimization, it is necessary to develop new methods to avoid these problems, and at the same time to enhance the performance in feature extraction and recognition tasks.

In this paper, motivated by the high order tensor-based discriminant analysis methods and the $L_{2,1}$-norm regression for jointly sparse FSSL methods, we propose a novel method called joint tensor feature analysis (JTFA) for sparse tensor feature selection and subspace learning (FSSL). Our original idea is to introduce the $L_{2,1}$-norm regression for jointly sparse discriminant feature selection and extraction from the columns and rows of the image matrix (or from each mode of the high order tensor data) so as to enhance the algorithm's performance. The detailed motivations for proposing JTFA are stated in Section III-B.

The main contributions of this paper are as follows.

1) We adopt the idea of jointly sparse feature selection and extraction from the tensor data to propose a new discriminant analysis method called JTFA.
2) We present a novel method on how to modify the classical discriminant analysis method in regression form for joint tensor feature extraction. Similarly, other regularizers based on different norms can be used in the same way for tensor regression.
3) The comprehensive analyses, including the convergent analysis, computational complexity analysis, and the essence of the proposed method/model, are also explored for the proposed JTFA.
4) Extensive experiments show that JTFA outperforms the classical subspace learning methods and the SDA for tensor objective recognition.

The rest of this paper is organized as follows. Section II briefly discusses the related work and the details of the proposed algorithm are shown in Section III. In Section IV, theoretical analyses, including the convergence, computational complexity, and the essence of the optimization model, are shown. Experiments are presented in Section V, and the conclusion is given in Section VI.

II. Related Work

In this section, we briefly discuss some related works on feature extraction and spare feature selection.

LDA is one of the classical discriminant feature extraction methods. Based on the tensor representation, the classical LDA was extended to 2-DLDA [10] and MLDA [11]. By using the differential criterion, Li *et al.* [14] proposed the MMC method, which was also extended to the multilinear cases, i.e., GTDA [18] and TMMC [19]. The common property of these methods is to use the eigen-decomposition method to compute the projections. With the development of the tensor learning methods, biologically inspired features [21], [22], and MMR [24] were recently proposed for objective recognition. However, these methods only have the function of dimensionality reduction and cannot perform sparse feature selection.

The L_1 norm-based sparse regularization methods provide an effective way for simultaneous feature extraction and sparse feature selection. The representative methods include SPCA [30], SDA [31], and its modifications [34]–[36]. These methods focus on the high-dimensional vector representations for sparse dimensionality reduction and feature selection. However, the nonzero elements of the projections derived by the L_1 norm regression can lie on any location of the learned projections and thus are absent for the joint sparsity.

The $L_{2,1}$-norm-based joint feature selection is a novel way for sparse feature selection. The jointly sparse learning methods such as FSSL and those proposed in [39]–[43] can obtain the jointly sparse subspace. The nonzero elements of the learned projections based on the $L_{2,1}$-norm regression lie on the same location and thus have the function of jointly sparse feature selection. However, these existing jointly sparse learning methods only focus on the high-dimensional vector and how to deal with the high-order tensor data and enhance the discriminant capability of the jointly sparse learning methods remains unsolved. To deal with this problem, the proposed JTFA inherits the discriminant capability from the multiLDA and the joint sparsity from the $L_{2,1}$-norm-based learning methods. Based on the novel definitions of the between-class and within-class scatter matrices, the discriminant projection learning procedures are converted to the $L_{2,1}$-norm multilinear regularized regression, which will be shown in the following section.

III. JTFA

In this section, we first briefly present some basic multilinear notations, definitions, and operations following the same way as in previous tensor learning methods [9], [11], [37]. Since we focus on the jointly sparse feature selection and extraction, we present the key idea and the related definitions first. In order to realize the goal of jointly sparse feature selection on the tensor data, the $L_{2,1}$ norm penalty terms are added to the objective function to regularize the projections used for feature extraction and selection. At last, an iterative method is proposed to solve the optimization problem.

A. Preparations

In this paper, we follow the similar definitions and notations as in [9], [11], and [37]. 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., $\mathbf{A}, \mathbf{B}, \mathbf{S}, \boldsymbol{\Phi}$, 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. The purpose of the tensor feature extraction is to obtain a set of sparse projection matrix $\{\mathbf{U}_i \in R^{d_i \times m_i}, d_i \le m_i, i = 1, 2, \ldots, n\}$ that map the original high-order tensor data into a low-order tensor space

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

In order to compute the series of projection matrices, we need the following definitions.

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 $< \mathcal{X}, \mathcal{Y} >= \sum_{i_1,=1,\ldots,i_n=1}^{m_1 \times m_2 \times \cdots \times m_n} \mathcal{X}_{i_1,\ldots,i_n} \mathcal{Y}_{i_1,\ldots,i_n}$. 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 $D(\mathcal{X}, \dagger) = \|\mathcal{X} - \mathcal{Y}\|$.

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

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).

For a given matrix $\mathbf{A} = [a_{ij}] \in R^{d \times m}$, we denote the ith row of $\mathbf{A}$ by $\mathbf{a}^i$. The Frobenius norm of matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_F = \sqrt{\sum\nolimits_{i=1}^{d} \sum_{j=1}^{m} a_{ij}^2} = \sqrt{\sum\nolimits_{i=1}^{d} \left\|\mathbf{a}^i\right\|_2^2}. \tag{2}$$

The $L_{2,1}$-norm of a matrix was first introduced in [16] as rotational invariant L_1-norm and also used for multitask learning and tensor factorization. It is defined as

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum\nolimits_{j=1}^{m} a_{ij}^2} = \sum_{i=1}^{d} \left\|\mathbf{a}^i\right\|_2. \tag{3}$$

With the above preparations, we begin to present our method in the following subsections.

B. Motivations and the Novel Definitions

As mentioned in previous sections, we aim to perform jointly sparse feature selection and extraction on the tensor data. For simplicity, we take the image matrix (two order tensor) as an example. Since the image matrix contains a lot of redundant information (i.e., the columns or rows are correlated), not all the pixels are helpful for the feature extraction and recognition task. We expect to jointly extract the discriminant information sparsely embedded in the image matrix in some rows or columns. In other words, we tend to find the optimal jointly sparse matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ (sparse in column) for feature extraction from the image matrix $\mathbf{X}$ so as to obtain the small size feature matrix $\mathbf{U}_1^T \mathbf{X} \mathbf{U}_2$. In order to obtain the jointly sparse matrices $\mathbf{U}_i$ *s*, the $L_{2,1}$-norm regularized term is appended to the objective function with respect to these projective matrices. We introduce the new variables $\mathbf{P}_k$ *s* in the model so that the model can be formulated into regression form. Thus, it is convenient to use the $L_{2,1}$-norm regularized regression method to compute the optimal solution of the proposed objective function. By using the idea of LDA/MMC, the proposed objective function also aims to minimize the modified within-class tensor scatter value and maximize the modified between-class tensor scatter value. Therefore, we need the following novel definitions.

For simplicity, we suppose that there are C classes and each class has N_w training samples in this paper.

Definition 4: The modified within-class tensor scatter value S_W is defined as

$$S_W = \sum_{j=1}^{C} \sum_{i=1,\mathcal{X}_i \in C_j}^{N_w} \left\|\mathcal{X}_i - \bar{\mathcal{X}}_j \times_1 \mathbf{U}_1 \mathbf{P}_1^T \times_2 \mathbf{U}_2 \mathbf{P}_2^T \cdots \times_n \mathbf{U}_n \mathbf{P}_n^T\right\|_F^2 \tag{4}$$

where $\bar{\mathcal{X}}_j$ denotes the mean value of the tensor samples in the jth class and C_j denotes the training sample set of the jth class.

Definition 5: The modified between-class tensor scatter matrix S_B is defined as

$$S_B = \sum_{j=1}^{C} N_w \left\|\bar{\mathcal{X}}_j - \bar{\mathcal{X}} \times_1 \mathbf{U}_1 \mathbf{P}_1^T \times_2 \mathbf{U}_2 \mathbf{P}_2^T \cdots \times_n \mathbf{U}_n \mathbf{P}_n^T\right\|_F^2 \tag{5}$$

where $\bar{\mathcal{X}}_j$ denote the mean value of the tensor samples of jth class and $\bar{\mathcal{X}}$ denote the mean value of all the training samples, respectively.

C. Objective Function of JTFA and its Solutions

The objective function of JTFA is to minimize the tensor discriminant function of the $L_{2,1}$-norm penalty optimization problem with a set of constraints

$$\begin{aligned}&\min J(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n, P_1, P_2, \ldots, \mathbf{P}_n)\\&\triangleq \min S_W - \mu S_B + \sum_{i=1}^{n} \gamma_i \|\mathbf{U}_i\|_{2,1}\\&\text{s.t. } \mathbf{P}_1^T\mathbf{P}_1 = \mathbf{I}_1, \mathbf{P}_2^T\mathbf{P}_2 = \mathbf{I}_2, \ldots, \mathbf{P}_n^T\mathbf{P}_n = \mathbf{I}_n \qquad (6)\end{aligned}$$

where μ is the parameter to balance the two scatter values and γ_i s are the parameters for the regularization terms.

For $i \neq k$, when all the $\mathbf{U}_i$ s and $\mathbf{P}_i$ s are given, we obtain the mode-k unfolding form of the optimization

$$\begin{aligned}&\min_{\mathbf{U}_k,\mathbf{P}_k} J(\mathbf{U}_k, \mathbf{P}_k) = \min_{\mathbf{U}_k,\mathbf{P}_k} \mathbf{S}_W^{(k)} - \mu \mathbf{S}_B^{(k)} + \sum_{i=1}^{n} \gamma_i \|\mathbf{U}_i\|_{2,1}\\&\text{s.t. } \mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}_i \qquad (7)\end{aligned}$$

where

$$\mathbf{S}_W^{(k)} = \sum_{j=1}^{C} \sum_{i=1, X_i^{(k)} \in C_j}^{N_w} \left\| \left(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}}_j^{(k)} \mathbf{U}_k \mathbf{P}_k^T \right) \right\|_F^2 \qquad (8)$$

$$\mathbf{S}_B^{(k)} = \sum_{j=1}^{C} N_w \left\| \bar{\mathbf{X}}_j^{(k)} - \bar{\mathbf{X}}^{(k)} \mathbf{U}_k \mathbf{P}_k^T \right\|_F^2 \qquad (9)$$

where $\bar{\mathbf{X}}_j^{(k)}$ denotes the mean value of the mode-k flattening of the tensor samples in the jth class and $\bar{\mathbf{X}}^{(k)}$ denotes the mean value of the mode-k flattening of the tensor samples of all the training samples, that is

$$\begin{aligned}\bar{\mathbf{X}}_j^{(k)} \Leftarrow_k &\ \bar{\mathcal{X}} \times_1 \mathbf{U}_1\mathbf{P}_1^T \cdots \times_{k-1} \mathbf{U}_{k-1}\mathbf{P}_{k-1}^T \times_{k+1} \mathbf{U}_{k+1}\mathbf{P}_{k+1}^T \cdots\\&\times_n \mathbf{U}_n\mathbf{P}_n^T\end{aligned}$$

and

$$\begin{aligned}\bar{\mathbf{X}}^{(k)} \Leftarrow_k &\ \bar{\mathcal{X}} \times_1 \mathbf{U}_1\mathbf{P}_1^T \cdots \times_{k-1} \mathbf{U}_{k-1}\mathbf{P}_{k-1}^T \times_{k+1} \mathbf{U}_{k+1}\mathbf{P}_{k+1}^T \cdots\\&\times_n \mathbf{U}_n\mathbf{P}_n^T.\end{aligned}$$

Since $\mathbf{U}_i$s are fixed except for any $i \neq k$, then $\sum_{i \neq k} \|\mathbf{U}_i\|_{2,1}$ becomes a constant. From (8) and (9) and using the constraint $\mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}_k$ we have

$$\begin{aligned}\mathbf{S}_W^{(k)} = tr\Bigg(&\sum_{j=1}^{C} \sum_{i=1, \mathbf{X}_i^{(k)} \in C_j}^{N_w} \mathbf{X}_i^{(k)T}\mathbf{X}_i^{(k)} - 2\mathbf{P}_k^T\mathbf{X}_i^{(k)T}\bar{\mathbf{X}}_j^{(k)}\mathbf{U}_k\\&+ \mathbf{U}_k^T\bar{\mathbf{X}}_j^{(k)T}\bar{\mathbf{X}}_j^{(k)}\mathbf{U}_k\Bigg)\end{aligned}$$

$$\begin{aligned}\mathbf{S}_B^{(k)} = tr\Bigg(&\sum_{j=1}^{C} \mu N_w \left(\bar{\mathbf{X}}_i^{(k)T}\bar{\mathbf{X}}_i^{(k)} - 2\mathbf{P}_k^T\bar{\mathbf{X}}_i^{(k)T}\bar{\mathbf{X}}^{(k)}\mathbf{U}_k\right.\\&+ \mathbf{U}_k^T\bar{\mathbf{X}}^{(k)T}\bar{\mathbf{X}}^{(k)}\mathbf{U}_k\Bigg).\end{aligned}$$

For ease of representation, we integrate the terms in $\mathbf{S}_W^{(k)}$ and $\mathbf{S}_B^{(k)}$ related to $\mathbf{P}_k$ and $\mathbf{U}_k$ together. Thus, we denote

$$\tilde{\mathbf{S}}_1^{(k)} = \sum_{j=1}^{C} \sum_{i=1, \mathbf{X}_i^{(k)} \in C_j}^{N_w} \mathbf{X}_i^{(k)T}\bar{\mathbf{X}}_j^{(k)} - \mu \sum_{i=1}^{C} N_w \bar{\mathbf{X}}_i^{(k)T}\bar{\mathbf{X}}^{(k)}. \qquad (10)$$

Similarly, we integrate the terms only related to the variable $\mathbf{U}_k$ together. Thus, we have

$$\begin{aligned}\tilde{\mathbf{S}}_2^{(k)} &= \sum_{j=1}^{C} \sum_{i=1,, \mathbf{X}_i^{(k)} \in C_j}^{N_w} \bar{\mathbf{X}}_j^{(k)T}\bar{\mathbf{X}}_j^{(k)} - \mu \sum_{i=1}^{C} N_w \bar{\mathbf{X}}^{(k)T}\bar{\mathbf{X}}^{(k)}\\&= \sum_{j=1}^{C} \sum_{i=1}^{N_w} \bar{\mathbf{X}}_j^{(k)T}\bar{\mathbf{X}}_j^{(k)} - \mu \sum_{i=1}^{C} N_w \bar{\mathbf{X}}^{(k)T}\bar{\mathbf{X}}^{(k)}. \qquad (11)\end{aligned}$$

When all the $\mathbf{U}_i$s and $\mathbf{P}_i$s are fixed except for any $i \neq k$, by discarding the constants in $\mathbf{S}_W^{(k)}$ and $\mathbf{S}_B^{(k)}$ and using (10) and (11) for representation, we have the following optimization problem from objective function (6):

$$\begin{aligned}&\min_{\mathbf{U}_k,\mathbf{P}_k} J(\mathbf{U}_k, \mathbf{P}_k)\\&= \min_{\mathbf{U}_k,\mathbf{P}_k} tr\left(-2\mathbf{P}_k^T\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k + \mathbf{U}_k^T\tilde{\mathbf{S}}_2^{(k)}\mathbf{U}_k\right) + \gamma_i \|\mathbf{U}_k\|_{2,1}\\&\text{s.t. } \mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}_k. \qquad (12)\end{aligned}$$

The above optimization problem is the core component of the proposed algorithm, which will be solved in next subsection.

D. Optimal Solution

From the analysis of the previous subsection, we obtain the mode-k minimization problem of the $L_{2,1}$-norm regularized optimization. According to the definition of the $L_{2,1}$-norm about the projection matrix $\mathbf{U}_k$, we could define a diagonal matrix $\mathbf{G}^k$ with the ith diagonal element as

$$\mathbf{G}_{ii}^k = \frac{1}{2\left\|\mathbf{u}_i^k\right\|_2} \qquad (13)$$

where $\mathbf{u}_i^k$ denotes the ith row of matrix $\mathbf{U}_k$. Thus the objective function in (12) is equivalent to

$$\begin{aligned}&\min_{\mathbf{U}_k,\mathbf{P}_k} J(\mathbf{U}_k, \mathbf{P}_k)\\&= \min_{\mathbf{P}_k,\mathbf{U}_k} tr\left(-2\mathbf{P}_k^T\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k + \mathbf{U}_k^T\left(\tilde{\mathbf{S}}_2^{(k)} + \gamma_i\mathbf{G}^k\right)\mathbf{U}_k\right)\\&\text{s.t. } \mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}_k. \qquad (14)\end{aligned}$$

For the above optimization problem, we design an iterative algorithm to solve it. The algorithm step can be stated as follows. First, fix $\mathbf{P}_k$ to compute $\mathbf{U}_k$, then update $\mathbf{G}^k$ and fix $\mathbf{U}_k$ to compute $\mathbf{P}_k$; iterate these two steps until the algorithm converges.

For the given $\mathbf{P}_k$, taking the partial deviation of (14) with respect to $\mathbf{U}_k$ to be equal to 0, we obtain

$$-\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k + \left(\tilde{\mathbf{S}}_2^{(k)} + \gamma_i\mathbf{G}^k\right)\mathbf{U}_k = 0.$$

This gives

$$\mathbf{U}_k = \left(\tilde{\mathbf{S}}_2^{(k)} + \gamma_i\mathbf{G}^k\right)^{-1}\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k. \qquad (15)$$

When $\mathbf{U}_k$ is given, we need to update the matrix $\mathbf{G}^k$ in (14). Then the minimum problem in (14) is equivalent to the maximum problem as follows:

$$\begin{aligned}&\max_{\mathbf{P}_k} tr\left(\mathbf{P}_k^T\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k\right)\\&\text{s.t.}\mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}_k. \qquad (16)\end{aligned}$$

TABLE I
JTFA 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 $T$, dimensions $d_k (d_k \le m_k)$
Output: Low-dimensional features $\mathcal{Y}_i \ (i = 1,2,...,N)$
Step 1: Compute the $\bar{\mathcal{X}}$, $\bar{\mathcal{X}}_i$ s, and initialize $\mathbf{G}_1^{(0)}, \mathbf{G}_2^{(0)}, \cdots, \mathbf{G}_n^{(0)}, \mathbf{P}_1^{(0)}, \mathbf{P}_2^{(0)}, \cdots, \mathbf{P}_n^{(0)}$.
Step 2: For $k = 1,2,...,n$
        Unfold the tensors to construct matrix $\tilde{\mathbf{S}}_1^{(k)}$ (and $\tilde{\mathbf{S}}_2^{(k)}$)
        For $t = 1:T$
            - Compute $\mathbf{U}_k$ using (15)
            - Compute $\mathbf{P}_k$ using (18)
            - Compute $\mathbf{G}^k$ using (13).
        End
    End
Step 4: Normalize each column vectors of $\mathbf{U}_k$ to be identity vectors and output the final matrix $\mathbf{U}_k$ for feature extraction.
Step 5: Project the samples onto the low-dimensional subspace to obtain $\mathcal{Y}_i = \mathcal{X}_i \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \cdots \times_n \mathbf{U}_n$.
```

According to [30, Th. 4], the optimal solution for the above problem is given by the SVD of $\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k$. Let

$$\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k = \hat{\mathbf{U}}_k\hat{\mathbf{D}}_k\hat{\mathbf{V}}_k^T. \tag{17}$$

Then the optimal solution of (16) is

$$\mathbf{P}_k = \hat{\mathbf{U}}_k\hat{\mathbf{V}}_k^T. \tag{18}$$

In fact, as computing the matrix $\mathbf{U}_k$ in (15) needs the input of $\mathbf{G}^k$ which is still not directly obtained. Therefore, we need to compute the matrix $\mathbf{G}^k$ in the designed iterative algorithm. Once the $\mathbf{U}_k$ and $\mathbf{P}_k$ are obtained, we can compute the other pair of variables in the same way. Iterating the above procedures will give the local optimal solutions of the algorithm. The algorithm details are described in Table I.

E. Comparison and Discussion

From the above four subsections, we can find the main differences between JTFA and the previous methods. FSSL is a high-dimensional vector-based method, which uses the spectral vectors and $L_{2,1}$-norm regression for feature selection. Although MMR [24] also introduced the concept of tensor, but MMR mainly focuses on the image matrix (i.e., second order tensor) and the label indicator matrix Y as the one in ridge regression is used for regression. Thus, the essence of MMR is the second order tensor extension of the ridge regression or least square regression. JTFA is significantly different from STL [23] in the objective function and the method used in computing the optimal solution since STL is the multilinear extension of the convex optimization-based learning algorithm.

There are some common properties of MLDA, TMMC, and JTFA. All of them directly use the idea of Fisher discriminant analysis and the tensor as the input. However, the projections of MLDA, TMMC, MMR, and STL are not (jointly) sparse. Thus they do not have the function of jointly sparse feature selection. Unlike the other tensor-based algorithms, JTFA can rewrite the multiLDA method into regression form and use $L_{2,1}$-norm regression to learn a set of jointly sparse projections. Thus, the advantages of JTFA against MLDA, TMMC, MMR, and STL are that JTFA not only can perform feature extraction but also jointly sparse feature selection from each mode of the tensor data.

In short, JTFA is a novel method on the tensor discriminant analysis, which provides a representative way on how to obtain the multilinear jointly sparse projections. The most significant difference between JFTA and the other tensor-based methods is that JTFA can obtain a set of jointly sparse projections for achieving the advantages of simultaneous feature extraction and feature selection from each mode of the tensors.

IV. THEORETICAL ANALYSIS

In this section, we will further give the theoretical analysis of the proposed algorithm, which includes the convergent analysis, computational complexity analysis, and the essence of the optimization method.

A. Convergence

In order to prove the convergence of the proposed algorithm, we need the following lemma.

Lemma 1 [37]: For any nonzero vectors $\mathbf{p}, \mathbf{p}_t \in R^c$, the following inequality holds:

$$\|\mathbf{p}\|_2 - \frac{\|\mathbf{p}\|_2^2}{2\,\|\mathbf{p}_t\|_2} \le \|\mathbf{p}_t\|_2 - \frac{\|\mathbf{p}_t\|_2^2}{2\,\|\mathbf{p}_t\|_2}. \tag{19}$$

With Lemma 1, we have the following theorem.

Theorem 1: Suppose all the variables in the objective function are given except for $\mathbf{U}_k$ and $\mathbf{P}_k$. The iteration approach presented in Section III-D will monotonically decrease the objective function $J(\mathbf{U}_k, \mathbf{P}_k)$ in each iteration and converge to the local optimum of the problem.

Proof: For ease of representation, we denote the objective function of (14) as $J(\mathbf{U}_k, \mathbf{P}_k) = J(\mathbf{U}_k, \mathbf{P}_k, \mathbf{G}^k)$. Suppose for the $t-1$th iteration, we obtain $\mathbf{U}_k^{(t-1)}$ and $\mathbf{P}_k^{(t-1)}$. From (15), we can find that

$$J\left(\mathbf{U}_k^{(t)}, \mathbf{P}_k^{(t-1)}, \mathbf{G}^{k,(t-1)}\right) \le J\left(\mathbf{U}_k^{(t-1)}, \mathbf{P}_k^{(t-1)}, \mathbf{G}^{k,(t-1)}\right). \tag{20}$$

Since the SVD gives the optimal $\mathbf{P}_k^{(t)}$ which further decreases the objective function, we have

$$J\left(\mathbf{U}_k^{(t)}, \mathbf{P}_k^{(t)}, \mathbf{G}^{k,(t-1)}\right) \leq J\left(\mathbf{U}_k^{(t-1)}, \mathbf{P}_k^{(t-1)}, \mathbf{G}^{k,(t-1)}\right). \tag{21}$$

Once the optimal $\mathbf{P}_k^{(t)}$ and $\mathbf{U}_k^{(t)}$ are obtained, we have

$$\begin{aligned} tr &\left(-2\mathbf{P}_k^{(t)T}\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k^{(t)} + \mathbf{U}_k^{(t)T}\left(\tilde{\mathbf{S}}_2^{(k)} + \gamma_i\mathbf{G}^{k(t-1)}\right)\mathbf{U}_k^{(t)}\right) \\ &\leq tr\left(-2\mathbf{P}_k^{(t-1)T}\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k^{(t-1)}\right. \\ &\quad \left. + \mathbf{U}_k^{(t-1)T}\left(\tilde{\mathbf{S}}_2^{(k)} + \gamma_i\mathbf{G}^{k(t-1)}\right)\mathbf{U}_k^{(t-1)}\right). \end{aligned} \tag{22}$$

That is

$$\begin{aligned} tr&\left(-2\mathbf{P}_k^{(t)T}\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k^{(t)} + \mathbf{U}_k^{(t)T}\tilde{\mathbf{S}}_2^{(k)}\mathbf{U}_k^{(t)}\right) + \gamma_k\sum_i\frac{\left\|\mathbf{u}_i^{k,(t)}\right\|_2^2}{2\left\|\mathbf{u}_i^{k,(t-1)}\right\|_2} \\ &\leq tr\left(-2\mathbf{P}_k^{(t-1)T}\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k^{(t-1)} + \mathbf{U}_k^{(t-1)T}\tilde{\mathbf{S}}_2^{(k)}\mathbf{U}_k^{(t-1)}\right) \\ &\quad + \gamma_k\sum_i\frac{\left\|\mathbf{u}_i^{k,(t-1)}\right\|_2^2}{2\left\|\mathbf{u}_i^{k,(t-1)}\right\|_2}. \end{aligned} \tag{23}$$

According to the Lemma 1, following the same way as in [37], it is easy to show that:

$$\begin{aligned} -\gamma_k&\left(\sum_i\left\|u_i^{k,(t)}\right\|_2 - \sum_i\frac{\left\|\mathbf{u}_i^{k,(t)}\right\|_2^2}{2\left\|\mathbf{u}^{i,(t-1)}\right\|_2}\right) \\ &\leq -\gamma_k\left(\sum_i\left\|\mathbf{u}_i^{k,(t-1)}\right\|_2 - \sum_i\frac{\left\|\mathbf{u}_i^{k,(t-1)}\right\|_2^2}{2\left\|\mathbf{u}_i^{k,(t-1)}\right\|_2}\right). \end{aligned} \tag{24}$$

Then combining (23) and (24), we obtain

$$\begin{aligned} tr&\left(-2\mathbf{P}_k^{(t)T}\tilde{S}_1^{(k)}\mathbf{U}_k^{(t)} + \mathbf{U}_k^{(t)T}\tilde{S}_2^{(k)}\mathbf{U}_k^{(t)}\right) + \gamma_k\left\|\mathbf{U}_k^{(t)}\right\|_{21} \\ &\leq tr\left(-2\mathbf{P}_k^{(t-1)T}\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k^{(t-1)} + \mathbf{U}_k^{(t-1)T}\tilde{\mathbf{S}}_2^{(k)}\mathbf{U}_k^{(t-1)}\right) \\ &\quad + \gamma_k\left\|\mathbf{U}_k^{(t-1)}\right\|_{21}. \end{aligned} \tag{25}$$

That is

$$\begin{aligned} J\left(\mathbf{U}_k^{(t)}, \mathbf{P}_k^{(t)}\right) &= J\left(\mathbf{U}_k^{(t)}, \mathbf{P}_k^{(t)}, G^{k,(t)}\right) \\ &\leq J\left(\mathbf{U}_k^{(t-1)}, \mathbf{P}_k^{(t-1)}, G^{k,(t-1)}\right) \\ &= J\left(\mathbf{U}_k^{(t-1)}, \mathbf{P}_k^{(t-1)}\right). \end{aligned} \tag{26}$$

Therefore, the algorithm will converge to the local optimum of the problem (12). ■

Theorem 2: The iterative algorithm will monotonically decrease the objective function $J(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n, \mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n)$ in each iteration and converge to the local optimum of the tensor problem.

Proof: Let $J(\mathbf{U}_1^{(t-1)}, \mathbf{U}_2^{(t-1)}, \ldots, \mathbf{U}_n^{(t-1)}, \mathbf{P}_1^{(t-1)}, \mathbf{P}_2^{(t-1)}, \ldots, \mathbf{P}_n^{(t-1)})$ be the objective function of the proposed method. We need to prove that it is nonincreasing and has a lower bound (at least bigger than a constant const $>$ 0). By frequently using Theorem 1, we can conclude that

$$\begin{aligned} J&\left(\mathbf{U}_1^{(t-1)}, \mathbf{U}_2^{(t-1)}, \ldots, \mathbf{U}_n^{(t-1)}, \mathbf{P}_1^{(t-1)}, \mathbf{P}_2^{(t-1)}, \ldots, \mathbf{P}_n^{(t-1)}\right) \\ &\geq J\left(\mathbf{U}_1^{(t)}, \mathbf{U}_2^{(t-1)}, \ldots, \mathbf{U}_n^{(t-1)}, \mathbf{P}_1^{(t)}, \mathbf{P}_2^{(t-1)}, \ldots, \mathbf{P}_n^{(t-1)}\right) \\ &\geq J\left(\mathbf{U}_1^{(t)}, \mathbf{U}_2^{(t)}, \ldots, \mathbf{U}_n^{(t-1)}, \mathbf{P}_1^{(t)}, \mathbf{P}_2^{(t)}, \ldots, \mathbf{P}_n^{(t-1)}\right) \geq \ldots \\ &\geq J\left(\mathbf{U}_1^{(t)}, \mathbf{U}_2^{(t)}, \ldots, \mathbf{U}_n^{(t)}, \mathbf{P}_1^{(t)}, \mathbf{P}_2^{(t)}, \ldots, \mathbf{P}_n^{(t)}\right) > \text{const}. \end{aligned}$$

Therefore, the objective function of (6) will converge to a local optimum. ■

B. Computational Complexity Analysis

For ease of understanding, we suppose that each mode of the tensor has the same size, i.e., $m_1 = m_2 = \cdots = m_n = m$, and the number of the training tensors is N. The main computational complexity of JTFA algorithm is to compute the scatter matrix $\tilde{\mathbf{S}}_1^{(k)}$, the matrix $\mathbf{U}_k$ in (15), the multilinear projection operation, and SVD of (17) in each step. The computation needed to compute the scatter matrix $\tilde{\mathbf{S}}_1^{(k)}$ is in the order of $O(Nnm^{n+1})$ (upper bounded). Computing the matrix $\mathbf{U}_k$ in (15) needs $O(m^3)$. Computing the multilinear projection needs $O(nm^{n+1})$. SVD of (17) also needs $O(m^3)$. If the algorithm needs T iteration steps, then the total computational complexity is in the order of $O(TNnm^{n+1} + Tnm^3 + Tnm^{n+1})$.

C. Intrinsic Connections Between $\mathbf{U}_k$ and $\mathbf{P}_k$

In this subsection, we explore the close relationship between the subspace $\mathbf{U}_k$ and $\mathbf{P}_k$.

Firstly, we have the following conclusion.

Lemma 2: $\tilde{\mathbf{S}}_1^{(k)} = \tilde{\mathbf{S}}_2^{(k)}$ for any k.

Proof: According to the definitions of $\bar{\mathbf{X}}_j^{(k)}$ and $\bar{\mathbf{X}}^{(k)}$, we have

$$\begin{aligned} \tilde{\mathbf{S}}_1^{(k)} &= \sum_{j=1}^{C}\sum_{i=1,\mathbf{X}_i^{(k)}\in C_j}^{N_w}\mathbf{X}_i^{(k)T}\bar{\mathbf{X}}_j^{(k)} - \mu\sum_{i=1}^{C}N_w\bar{\mathbf{X}}_i^{(k)T}\bar{\mathbf{X}}^{(k)} \\ &= \sum_{j=1}^{C}N_w\left(\sum_{i=1,\mathbf{X}_i^{(k)}\in C_j}^{N_w}\frac{1}{N_w}\mathbf{X}_i^{(k)T}\right)\bar{\mathbf{X}}_j^{(k)} \\ &\quad - \mu\left(\sum_{i=1}^{C}N_w\bar{\mathbf{X}}_i^{(k)T}\right)\bar{\mathbf{X}}^{(k)} \\ &= \sum_{j=1}^{C}N_w\bar{\mathbf{X}}_j^{(k)T}\bar{\mathbf{X}}_j^{(k)} - \mu N_w C\bar{\mathbf{X}}^{(k)T}\bar{\mathbf{X}}^{(k)} \\ &= \sum_{j=1}^{C}\sum_{i=1}^{N_w}\bar{\mathbf{X}}_j^{(k)T}\bar{\mathbf{X}}_j^{(k)} - \mu\sum_{i=1}^{C}N_w\bar{\mathbf{X}}^{(k)T}\bar{\mathbf{X}}^{(k)} = \tilde{\mathbf{S}}_2^{(k)}. \end{aligned}$$

■

Lemma 2 indicates that even if the numerators are presented in different forms, but they are equal to each other. Since $\tilde{\mathbf{S}}_1^{(k)} = \tilde{\mathbf{S}}_2^{(k)}$, this property will decrease the computational burden in computing the scatter matrix. By using this property, we can prove the following theorem.

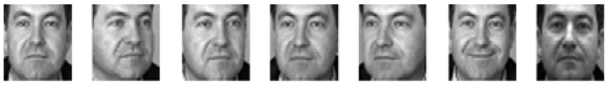


Fig. 1. Sample images of one person on FERET face database.

Theorem 3: If $\gamma_k \to 0(\forall k)$, then $\mathbf{U}_k \to \mathbf{P}_k(\forall k)$. If $\gamma_k = 0(\forall k)$, then $\mathbf{U}_k = \mathbf{P}_k$.

Proof: From (15), we know that $\mathbf{U}_k = (\tilde{\mathbf{S}}_2^{(k)} + \gamma_k \mathbf{G}^k)^{-1}\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k$. Using Lemma 2, we have $\mathbf{U}_k = (\tilde{\mathbf{S}}_1^{(k)} + \gamma_i \mathbf{G}^k)^{-1}\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k$. Then when $\gamma_k \to 0$, $\mathbf{U}_k \to (\tilde{\mathbf{S}}_1^{(k)} + 0G^k)^{-1}\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k = \mathbf{P}_k$. It is obvious that if $\gamma_k = 0(\forall k)$, then $\mathbf{U}_k = \mathbf{P}_k$. ■

D. *Essence of the Optimization Problem for* $\mathbf{U}_k$ *and* $\mathbf{P}_k$

In this subsection, we explore the essence of $\mathbf{U}_k$ and $\mathbf{P}_k$ in the optimization problem (14). Substituting (15) into (14) gives the following optimization problem:

$$\min_{\mathbf{P}_k} tr\left(-2\mathbf{P}_k^T\tilde{\mathbf{S}}_1^{(k)}\left(\tilde{\mathbf{S}}_1^{(k)} + \gamma_i\mathbf{G}^k\right)^{-1}\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k\right)$$
$$\text{s.t. } \mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}_k. \tag{27}$$

Obviously, the above minimization problem is equivalent to the following maximum problem:

$$\max_{\mathbf{P}_k} tr\left(\mathbf{P}_k^T\tilde{\mathbf{S}}_1^{(k)}\left(\tilde{\mathbf{S}}_1^{(k)} + \gamma_i\mathbf{G}^k\right)^{-1}\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k\right)$$
$$\text{s.t. } \mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}_k. \tag{28}$$

It is clear that the optimal solution is the standard eigen decomposition of the following eigen equation:

$$\tilde{\mathbf{S}}_1^{(k)}\left(\tilde{\mathbf{S}}_1^{(k)} + \gamma_i\mathbf{G}^k\right)^{-1}\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k = \mathbf{P}_k\mathbf{\Lambda} \tag{29}$$

where $\mathbf{\Lambda}$ is the eigenvalue matrix. Therefore $\mathbf{P}_k$ contains the eigenvectors corresponding to the larger eigenvalues of (29).

If $\mathbf{P}_k$ contains all the eigenvectors of (29), since $\mathbf{U}_k = (\tilde{\mathbf{S}}_1^{(k)} + \gamma_k\mathbf{G}^k)^{-1}\tilde{\mathbf{S}}_1^{(k)}\mathbf{P}_k$, from (29) we have

$$\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k = \mathbf{P}_k\mathbf{\Lambda} \Rightarrow \mathbf{P}_k^T\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k = \mathbf{\Lambda}\left(\text{since } \ \mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}_k\right). \tag{30}$$

From the above analysis, we can obtain the following interesting conclusion. The essence of the optimization problem are as follows. If $\gamma_k > 0$, $\mathbf{U}_k$is sparse in row, (15) indicates that the optimization problem of the proposed model (14) is to find an row-sparse $\mathbf{U}_k$ and an orthogonal matrix to diagonalize $\tilde{\mathbf{S}}_1^{(k)}$; If $\gamma_k = 0$ (or $\gamma_k \to 0$), $\mathbf{U}_k$ is not sparse and $\mathbf{U}_k = \mathbf{P}_k$ (or $\mathbf{U}_k \to \mathbf{P}_k$). That is, model (14) mainly finds the optimal nonsparse matrix to diagonalize scatter matrix $\tilde{\mathbf{S}}_1^{(k)}$, i.e., $\mathbf{P}_k^T\tilde{S}_1^{(k)}\mathbf{P}_k = \mathbf{\Lambda}$ (or $\mathbf{P}_k^T\tilde{\mathbf{S}}_1^{(k)}\mathbf{U}_k \to \mathbf{\Lambda}$). This is very similar to SVD or standard eigen decomposition on the scatter matrix, in other words, it is similar to the well-known PCA or MMC and its extensions, in which the optimal solution can be obtained by SVD or standard eigen decomposition.

V. Experiments

In this section, a set of experiments is presented to evaluate the proposed JTFA algorithm for tensor feature extraction and recognition. We compared it with the traditional LDA, the recently proposed L_1 norm-based sparse subspace learning methods SLDA [34], and the most related $L_{2,1}$-norm-based FSSL [38]. For FSSL, LDA graph was used and thus $c - 1$ projections were obtained. The tensor-based methods, i.e., MLDA [11] and TMMC [19], were also compared. The nearest neighborhood classifier with the Euclidean distance was used in all the experiments.

A. *Experiments on FERET Face Database*

The FERET face database is a result of the FERET program, which was sponsored by the U.S. 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 included 1400 images of 200 individuals (each individual has seven images) and involved 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 were resized to 40 × 40 pixels. The sample images of one person are shown in Fig. 1.

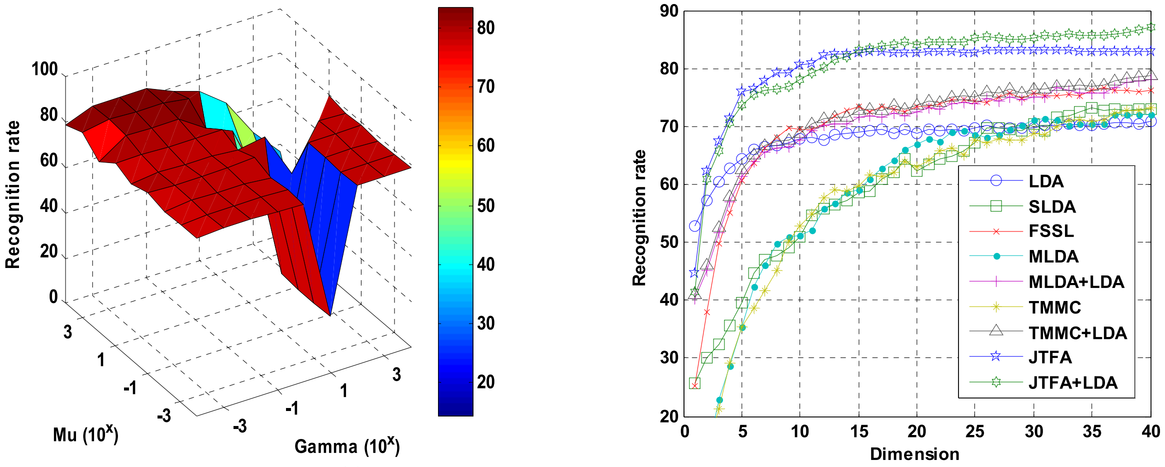
In this experiment, we investigate the performance of the recognition rates of different methods and the properties of the parameters of JTFA are also shown. In the experiments, four images were selected as the training set and the remaining three images were used for testing. We compared our methods, i.e., JTFA and JTFA + LDA, with the classical LDA method, sparse subspace methods, i.e., SLDA and FSSL, and tensor learning methods MLDA, TMMC, and MLDA + LDA and TMMC + LDA. Table II lists the recognition rates of each method. Fig. 2(a) shows the recognition rate versus the parameters γ and μ on the FERET face database. Fig. 2(b) shows the recognition rates versus the variations of the dimensions. It is noted that, for LDA and *+LDA (where * denotes MLDA, TMMC, and JTFA), the real dimension is the five times of the number marked in the horizontal axis.

From Fig. 2(a), we can see that the optimal value of the parameter μ lies on the area of $[10^1, 10^3]$ and the optimal γ lies on the area of $[10^{-3}, 10^1]$. In other words, when the parameters lie on these areas, JTFA is very robust to the parameters' variation. When the $\gamma \le 10^{-4}$ (i.e., $\gamma \to 0$), the recognition rate will be decreased, which indicates that the regularization parameter γ is very important for JTFA to achieve its best performance. The above empirical study will lead us to effectively use the JTFA algorithm so as to obtain good performance in the following experiments from Sections V-B to V-E.

From Table II and Fig. 2(b), it can be found that JTFA performs best among the compared methods, and LDA can enhance the performance of JTFA in image feature extraction.

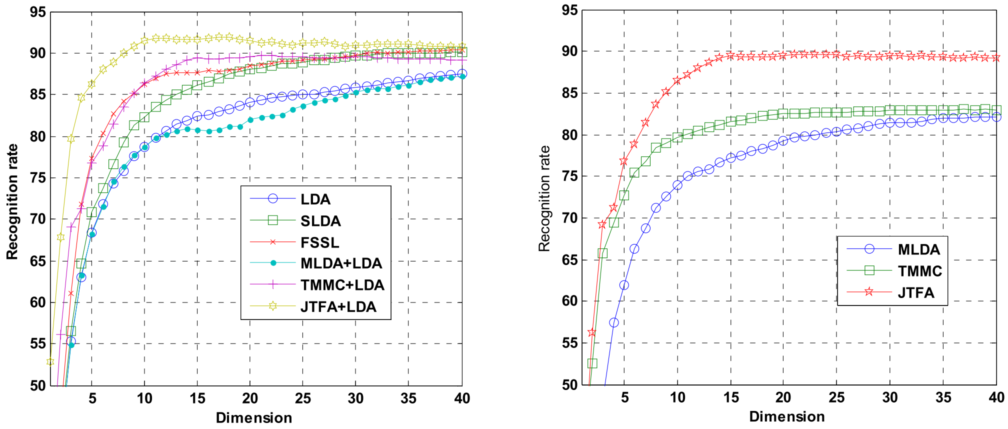
B. *Experiments on AR Face Database*

The AR face database [45] contains over 4000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions, and occlusions. The pictures of 120 individuals



(a) (b)

Fig. 2. (a) Recognition rate versus the parameters Gamma (γ) and Mu (μ) on the FERET face database. (b) Recognition rates (%) versus the dimensions of different methods on the FERET face databases.



(a) (b)

Fig. 3. Average recognition rates (%) versus the (a) dimensions of different vector-based discriminant analysis methods and the (b) tensor-based methods on the AR face database.

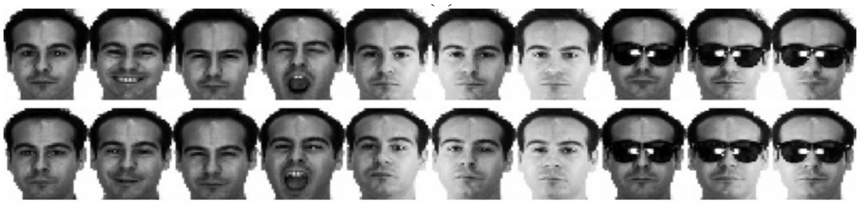


Fig. 4. Sample images of one person form the AR face database.

TABLE II
PERFORMANCE (RECOGNITION RATE AND DIMENSION) OF DIFFERENT METHODS ON FERET FACE DATABASE

methods	LDA	SLDA	FSSL	MLDA	MLDA +LDA	TMMC	TMMC +LDA	JTFA	JTFA +LDA
Recognition rate	70.83	73.16	76.67	72.00	78.00	73.17	78.67	83.17	87.17
Dimension	119	119	119	35×30	119	40×25	119	26×22	119

(65 men and 55 women) were selected and divided into two sessions (separated by two weeks) and each session contains 13 color images. Twenty images of these 120 individuals were selected and used in our experiments. The face portion of each image was manually cropped and then normalized to 50 × 40 pixels. The sample images of one person are shown in Fig. 2. These images vary as follows: 1) neutral expression; 2) smiling; 3) angry; 4) screaming; 5) left light on; 6) right

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

Training samples	LDA	SLDA	JSSL	MLDA	MLDA +LDA	TMMC	TMMC +LDA	JTFA	JTFA +LDA
3	81.68 119 ±7.20	82.17 119 ±6.73	81.80 119 ±9.09	73.71 40×38 ±9.20	81.21 119 ±12.11	75.11 36×38 ±6.09	82.22 119 ±9.02	80.33 34×30 ±7.81	83.69 119 ±4.53
5	87.53 119 ±3.29	90.12 119 ±5.57	90.43 119 ±7.15	82.18 40×38 ±10.16	87.24 119 ±11.08	83.00 39×38 ±6.55	89.67 119 ±9.65	87.06 32×30 ±5.01	91.91 119 ±4.48

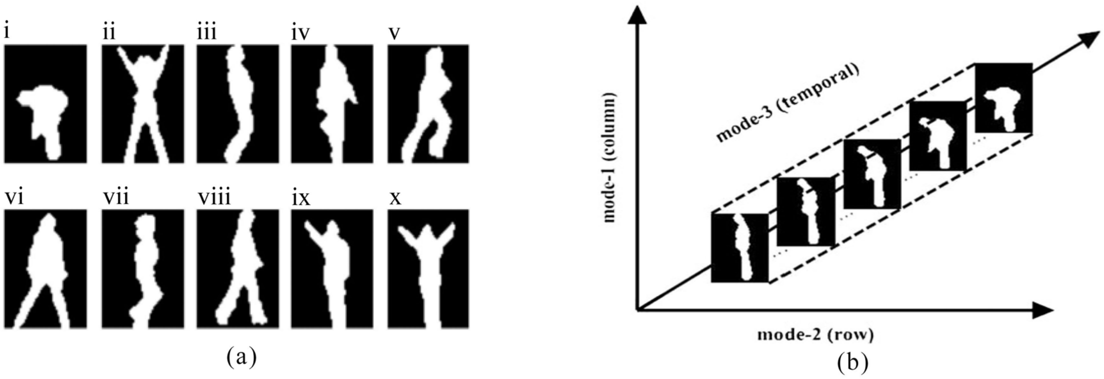


Fig. 5. (a) Key silhouettes of ten actions from the Weizmann database. (i) Bend. (ii) Jack. (iii) Jump. (iv) Pjump. (v) Run. (vi) Side. (vii) Skip. (viii) Walk. (ix) Wave1. (x) Wave2. (b) Example of the bending action in spatiotemporal domain from Weizmann database.

light on; 7) all sides light on; 8) wearing sun glasses; 9) wearing sun glasses and left light on; and 10) wearing sun glasses and right light on.

In the experiments, l ($l = 3, 5$) images of each individual were randomly selected and used for training, and half of the remaining images were used for validation and test, respectively. The optimal value of parameter μ was selected from the set $\{10^1, 10^2, 10^3\}$ and γ from the set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1\}$ since our empirical study indicates that JTFA can achieve the best performance on the different databases. The dimensions of different mode are varied from $[1, 40]$ for the tensor methods and from 5 to 200 with step as 5. For FSSL and SLDA, the regularization parameters were also selected from $\{10^{-3}, 10^{-2}, \ldots, 10^3\}$. The optimal parameters determined by the validation set were used to train the algorithm and the learned projections were used for feature extraction. The experiments were independently performed ten times and the average recognition rates on the test set are calculated and reported.

The performance of different methods on the AR face database are listed in Table III. The recognition rates versus the dimensions of vector-based discriminant analysis methods are shown in Fig. 3(a) and the tensor-based methods are shown in Fig. 3(b). It is noted that, for LDA and *+LDA (where * denotes MLDA, TMMC, and JTFA), the real dimension is the three times of the number marked in the horizontal axis.

C. Experiments on Weizmann Action Database

The experiment was performed on the Weizmann database [46], which is a commonly used database for human action recognition. There are 90 videos from ten 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. 5(a).

In order to represent the spatiotemporal feature of the samples, ten successive frames of each action were used to extract the temporal feature. Fig. 5(b) 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 × 24 × 10 pixels. It should be noted 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 aims to get high recognition accuracy, the methods used for feature extraction should be robust to starting frames and actions' variations. In the experiments, l ($l = 3, 5$) 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 V-B. The recognition rates of each method are listed in Table IV, and the variations of the average recognition rates versus the dimensions are shown in Fig. 6. It can be found that JTFA also outperforms the other algorithms in action tensor feature extraction.

D. Experiments on Cambridge Hand Gesture Database

The Cambridge hand gesture database [47] consists of 900 image sequences of nine gesture classes, which are defined by three primitive hand shapes and motions. The objective of

TABLE IV
PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION, AND DIMENSION) OF DIFFERENT METHODS ON THE WEIZMANN ACTION DATABASE

Training samples/object	Method	LDA	SLDA	FSSL	MLDA	MLDA +LDA	TMMC	TMMC +LDA	JTFA	JTFA +LDA
3	Recognition rate	64.22	64.46	66.44	64.04	65.12	64.32	65.66	67.88	69.41
	Standard deviation	±5.39	±4.92	±6.24	±6.08	±5.43	±3.54	±5.70	±3.88	±3.42
	Dimension	9	27	9	10^3	9	8^3	9	10^3	9
5	Recognition rate	73.27	74.06	75.92	75.16	74.74	75.21	74.20	77.67	78.58
	Standard deviation	3.18	±4.41	±4.56	±3.94	±3.88	±3.86	±3.23	±3.52	3.85
	Dimension	9	28	9	10^3	9	8^3	9	10^3	9

TABLE V
PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION, AND DIMENSION) OF DIFFERENT METHODS ON THE CAMBRIDGE HAND GESTURE DATABASE

Method	LDA	SLDA	FSSL	MLDA	MLDA +LDA	TMMC	TMMC +LDA	JTFA	JTFA +LDA
Recognition rate	75.80	76.05	77.63	78.86	79.28	80.68	80.90	83.59	84.00
Standard deviation	±2.17	±2.42	±2.00	±1.63	±3.94	±1.91	±2.17	±2.57	±2.58
Dimension	9	18	8	4^3	9	6^3	9	6^3	9

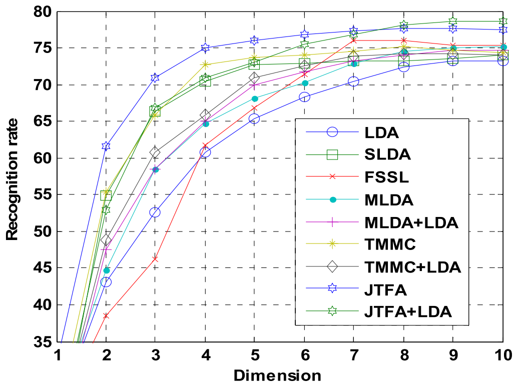


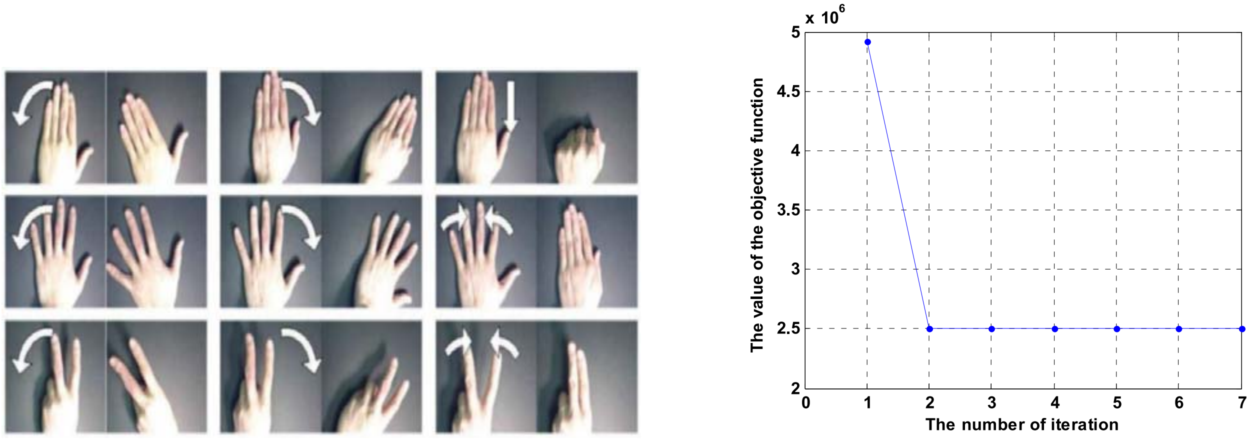
Fig. 6. Average recognition rates (%) versus the dimensions of different methods on the Weizmann action database.

using this data set is to classify different shapes as well as different motions at a time. Each class contains 100 image sequences (five different illuminations × ten arbitrary motions × two subjects). Each sequence was recorded in front of a fixed camera having roughly isolated gestures in space and time. Thus, fairly large intraclass variations in spatial and temporal alignment are reflected in the data set. Some sample images of nine different gesture classes are shown in Fig. 7(a). The experimental procedures are the same as in Weizmann action database. The recognition rates of each method are listed in Table V.

E. Experimental Results and Discussions

From the experimental results listed in Tables II–V and the figures presented in previous subsections, we have the following observations and corresponding analyses.

1) JTFA performs better than the previous tensor learning algorithms such as MLDA and TMMC. JTFA + LDA performs better than JTFA and LDA. Among all the results listed in this paper, JTFA + LDA obtains the best performance. This indicates that the selected features by JTFA can further enhance the LDAs performance.
2) For the LDA methods, SLDA and FSSL usually perform better than the classical LDA in image or tensor feature extraction. Thus, feature selection is very important for enhancing the recognition rates.
3) The image matrix-based methods (i.e., second order tensor-based MLDA, TMMC, and JTFA) may not be superior to the vector-based methods such as LDA, SLDA, and FSSL. This can be found from the experimental results on AR face database listed on Table III. However, with the increasing number of the order (i.e., when the number of the order is 3), the tensor-based methods concisely perform better than the vector-based methods. The possible reason is that with the rapid growth of the dimensions, the over-fitting of these methods and the decrease of the computational accuracy in computing the optimal solutions will possibly lead to deteriorate the performance.
4) Introducing the regularization term for sparse feature selection usually can enhance the performance. For example, SLDA performs better than LDA, and JTFA performs better than MMC, although they use very similar idea in feature extraction. Moreover, the two stage strategy of "*+LDA" (where * denotes the tensor learning methods) can further improve the algorithms' performance.
5) Although there are different expressions, lighting conditions and the starting frame of the tensor, JTFA is more robust to the other tensor-based discriminant analysis methods. The key reason is that JTFA has the sparse feature selection capability, which reduces the negative



(a)

(b)

Fig. 7. (a) Some sample images on Cambridge hand gesture database. (b) Example of the convergence curve of JTFA algorithm.

influences on the data obtained from the outside condition. This also demonstrates that the methods with the sparse constraint on the projections can obtain more robustness/stability.

6) Theoretical analysis in previous section indicates that JTFA is convergent. Fig. 7(b) shows an example on the convergence curve of JTFA algorithm on FERET face database. It can be seen that JTFA can converge within several iterations. There is similar property on other databases.

VI. Conclusion

In this paper, a novel tensor-based discriminant feature selection method called JTFA is proposed for sparse subspace learning. The $L_{2,1}$ norm was introduced in the JTFA model, and an iterative algorithm was designed to solve the optimization model. We proved the convergence of the proposed algorithm, and the theoretical analyses show that the designed iterative algorithm for the optimization problem converges to local optimum. Moreover, the computational complexity and the essence of the optimization procedures were also presented. We show that the essence of the optimization problem is to compute two matrices so as to diagonalize the scatter matrix similar to SVD or eigen-decomposition with sparse manner. Experiments on four well-known object recognition datasets showed that JTFA performed better than the traditional LDA methods and the miltilinear discriminant analysis methods. It can also be found from the experiments that the discriminant feature selection capability of the proposed JTFA is superior to the recently proposed L_1 norm and $L_{2,1}$ norm-based sparse discriminant analysis methods.

References

[1] 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.

[2] M. Turk, “Eigenfaces for recognition,” *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[3] 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.

[4] 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.

[5] 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, Jan. 2004.

[6] 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.

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

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

[9] H. Lu, K. N. 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.

[10] J. Ye, R. Janardan, and Q. Li, “Two-dimensional linear discriminant analysis,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17. Vancouver, BC, Canada, pp. 1569–1576, Jul. 2004.

[11] S. Yan *et al.*, “Multilinear discriminant analysis for face recognition,” *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 212–220, Jan. 2007.

[12] S. Wang, J. Yang, N. Zhang, and C. Zhou, “Tensor discriminant color space for face recognition,” *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2490–2501, Sep. 2011.

[13] S.-J. Wang *et al.*, “Sparse tensor discriminant color space for face verification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 876–888, Jun. 2012.

[14] 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.

[15] L. Zhang, L. Wang, and W. Lin, “Generalized biased discriminant analysis for content-based image retrieval,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 282–290, Feb. 2012.

[16] W. Yang and D. Dai, “Two-dimensional maximun margin feature extraction for face recognition,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 4, pp. 1002–1012, Aug. 2009.

[17] W. Yang, J. Wang, M. Ren, and J. Yang, “Feature extraction based on Laplacian bidirectional maximum margin criterion,” *Pattern Recognit.*, vol. 42, no. 11, pp. 2327–2344, 2009.

[18] 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. 1700–1715, Oct. 2007.

[19] 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.

[20] Y. Mu, W. Ding, M. Morabito, and D. Tao, “Empirical discriminative tensor analysis for crime forecasting,” in *Knowledge Science, Engineering and Management* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2011, pp. 293–304.

[21] Y. Mu and D. Tao, “Biologically inspired feature manifold for gait recognition,” *Neurocomputing*, vol. 73, nos. 4–6, pp. 895–902, Jan. 2010.

[22] Y. Mu, D. Tao, X. Li, and F. Murtagh, “Biologically inspired tensor features,” *Cogn. Comput.*, vol. 1, no. 4, pp. 327–341, Nov. 2009.

[23] D. Tao, X. Li, X. Wu, W. Hu, and S. J. Maybank, “Supervised tensor learning,” *Knowl. Inf. Syst.*, vol. 13, no. 1, pp. 1–42, Jan. 2007.

[24] Y. Su, X. Gao, X. Li, and D. Tao, “Multivariate multilinear regression,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1560–1573, Dec. 2012.

[25] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, “The collinearity problem in linear regression: The partial least squares approach to generalized inverses,” *SIAM J. Sci. Stat. Comput.*, vol. 5, no. 3, pp. 735–743, 1984.

[26] 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.

[27] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *J. Roy. Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[28] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[29] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J. Roy. Stat. Soc. B*, vol. 67, no. 2, pp. 301–320, 2005.

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

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

[32] A. Aspremont, F. Bach, I. Willow, and L. El Ghaoui, “Optimal solutions for sparse principal component analysis,” *J. Mach. Learn. Res.*, vol. 9, pp. 1269–1294, Jul. 2008.

[33] M. Journee, Y. Nesterov, P. Richtarik, and R. Sepulchre, “Generalized power method for sparse principal component analysis,” *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, Jan. 2010.

[34] 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, 2009.

[35] M. K. Ng, L. Liao, and L. Zhang, “On sparse linear discriminant analysis algorithm for high-dimensional data classification,” *Numer. Linear Algebr. Appl.*, vol. 18, no. 2, pp. 223–235, 2010.

[36] H. Zhao and S. Sun, “Sparse tensor embedding based multispectral face recognition,” *Neurocomputing*, vol. 133, no. 10, pp. 427–436, 2014.

[37] F. Nie, H. Huang, X. Cai, and C. Ding, “Efficient and robust feature selection via joint l2, 1 norms minimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23. Vancouver, BC, Canada, 2010, pp. 1813–1821.

[38] Q. Gu, Z. Li, and J. Han, “Joint feature selection and subspace learning,” in *Proc. 22nd Int. Joint Conf. Artif. Intell. (IJCAI)*, Barcelona, Spain, 2011, pp. 1294–1299.

[39] C. Hou, F. Nie, D. Yi, and Y. Wu, “Feature selection via joint embedding learing and sparse regression,” in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, Barcelona, Spain, 2011, pp. 1324–1329.

[40] Z. Ma *et al.*, “Discriminating joint feature analysis for multimedia data understanding,” *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.

[41] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, “Web image annotation via subspace-sparsity collaborated feature selection,” *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.

[42] C.-X. Ren, D.-Q. Dai, and H. Yan, “Robust classification using $\ell 2,1$-norm based regression model,” *Pattern Recognit.*, vol. 45, no. 7, pp. 2708–2718, Jul. 2012.

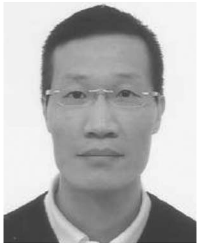
[43] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, “Joint embedding learning and sparse regression: A framework for unsupervised feature selection,” *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[44] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET evaluation methodology for face recognition algorithms,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[45] A. A. Martinez and R. Benavente, “The AR face database,” Comput. Vis. Center, Barcelona, Spain, CVC Tech. Rep. #24, 1998.

[46] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[47] T.-K. Kin, K.-Y. K. Wong, and R. Cipolla, “Tensor canonical correlation analysis for action classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, 2007, pp. 1–8.



Wai Keung Wong received the Ph.D. degree from Hong Kong Polytechnic University, Hong Kong.

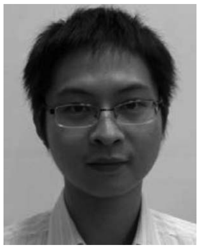
He is currently a Professor with Hong Kong Polytechnic University. His current research interests include artificial intelligence, pattern recognition, and optimization of manufacturing scheduling, planning, and control. He has published over 50 scientific articles in referred journals, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, the *International Journal of Production Economics*, the *European Journal of Operational Research*, the *International Journal of Production Research*, *Computers in Industry*, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, and so on.

Zhihui Lai received the B.S. degree in mathematics from South China Normal University, Guangzhou, China, the M.S degree from Jinan University, Guangzhou, 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.

From 2010 to 2014, he has been a Research Associate, a Post-Doctoral Fellow, and a Research Fellow with the Hong Kong Polytechnic University, Hong Kong. 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, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 1994, 1997, and 2005, respectively.

From 2005 to 2007, he was with Shenzhen Graduate School, Harbin Institute of Technology (HIT), Shenzhen, China, as a Post-Doctoral Research Fellow. He is currently a Professor with Shenzhen Graduate School, HIT. He was a Research Assistant Researcher with Hong Kong Polytechnic University, Hong Kong, from 2007 to 2008. His current research interests include pattern recognition, biometrics, and machine learning. He has published over 40 scientific papers.



Jiajun Wen received the M.S. degree in applied computer technology from the Guangdong University of Technology, Guangzhou, China, in 2010. He is currently pursuing the Ph.D. degree in computer science and technology from Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China.

He has been a Research Associate with Hong Kong Polytechnic University, Hong Kong, since 2013. His current research interests include pattern recognition and video analysis.

Chu Po Ho received the Ph.D. degree from Hong Kong Polytechnic University, Hong Kong.

He is currently an Assistant Professor with Hong Kong Polytechnic University. His current research interests include functional clothing design, pattern engineering, artificial intelligence, and optimization of manufacturing scheduling, planning, and control.