Multilinear Sparse Principal Component Analysis

Zhihui Lai, Yong Xu, Qingcai Chen, Jian Yang, *Member, IEEE*, and David Zhang, *Fellow, IEEE*

***Abstract*— In this brief, multilinear sparse principal component analysis (MSPCA) is proposed for feature extraction from the tensor data. MSPCA can be viewed as a further extension of the classical principal component analysis (PCA), sparse PCA (SPCA) and the recently proposed multilinear PCA (MPCA). The key operation of MSPCA is to rewrite the MPCA into multilinear regression forms and relax it for sparse regression. Differing from the recently proposed MPCA, MSPCA inherits the sparsity from the SPCA and iteratively learns a series of sparse projections that capture most of the variation of the tensor data. Each nonzero element in the sparse projections is selected from the most important variables/factors using the elastic net. Extensive experiments on Yale, Face Recognition Technology face databases, and COIL-20 object database encoded the object images as second-order tensors, and Weizmann action database as third-order tensors demonstrate that the proposed MSPCA algorithm has the potential to outperform the existing PCA-based subspace learning algorithms.**

***Index Terms*— Dimensionality reduction, face recognition, feature extraction, principal component analysis (PCA), sparse projections.**

I. Introduction

Principal component analysis (PCA) is one of the most widely used data preprocessing and feature extraction methods in the fields of computer vision and pattern recognition. As the classical unsupervised linear dimensionality reduction technique, PCA aims to obtain the most compact representations of the high-dimensional data under the sense of least square reconstruction error. Sirovich and Kirby [1], [2] used PCA to represent human faces for the first time. Therefore, Turk and Pentland [3] proposed a famous PCA-based face recognition method called eigenfaces.

To overcome the small sample size problem in PCA, Yang *et al.* [4] proposed the well-known 2-D PCA (2-DPCA) for directly performing feature extraction from the image matrices. Based on the same idea of 2-DPCA, many 2-D-based feature extraction methods [5]–[11] have been proposed for dimensionality reduction. Among these methods, generalized low rand approximations of matrices [7], which can be viewed as the special form of the concurrent subspaces analysis (CSA) [11], is the most representative method for 2-D image feature extraction. Recently, there is a great interest in higher order tensor analysis for feature extraction and recognition, and higher order tensor decomposition [12]–[14] has become popular in computer vision and pattern recognition [15]–[18]. More recently, multilinear PCA (MPCA) [19] and its uncorrelated variation [20] were also proposed for feature extraction on tensor objects. Zhao *et al.* [21] presented a framework that brings kernel methods and tensor decomposition techniques together such that nonlinear kernel-based strategy can be applied to tensor decompositions.

The PCA-based methods mentioned above all use L_2 norm as the measurement. However, recent research shows that introducing the L_1 norm for sparse feature selection not only can enhance the prediction accuracy, but also strengthen the generalization ability and the robustness for classification [22]–[25]. In [22], Tibshirani proposed the least absolute shrinkage and selection operator (LASSO) using the L_1 norm penalty for feature selection and obtained better performance than the ordinary least squares regression. The elastic net [24] generalizes the LASSO by combining both the ridge and lasso penalties and obtains better prediction accuracy. Using the elastic net in the regression-type optimization problem derived from PCA, sparse PCA (SPCA) [26] was proposed to compute the sparse principal component vectors for feature extraction and factor analysis.

However, SPCA can only deal with data expressed in the form of 1-D vector, and there exists much data such as image objects and action video that are intrinsically in the form of second or higher order tensors. Therefore, on the one hand, it is necessary to extend vector-based SPCA to second or higher order tensors so as to process any-order tensor data for dimensionality reduction using the most important variables/factors. On the other hand, when it is extended into second or higher order tensors, the efficiency and effectiveness of SPCA can be greatly improved since the extended method can directly operate on the tensor data. To these ends, the focus of this brief is on unsupervised subspace learning of the tensor data with sparseness constraint, i.e., SPCA on tensor analysis.

In this brief, multilinear SPCA (MSPCA) is proposed based on the following experimental/experiential observations. First, recent studies [4]–[7], [20], [27], [28] have shown that when the 1-D-based methods were extended to high-order tensors, the corresponding extension methods usually outperform the original ones, particularly in the cases of small sample sizes. Second, introducing the L_1 norm for sparse feature selection can enhance the prediction accuracy and strengthen the generalization ability and the robustness for classification [22]–[24], [25]. Thus, extending the PCA into higher order tensor

Manuscript received October 6, 2012; revised August 29, 2013; accepted December 21, 2013. Date of publication January 22, 2014; date of current version September 15, 2014. This work was supported in part by the Natural Science Foundation of China under Grant 61203376, Grant 61375012, Grant 61203247, Grant 61005005, Grant 61071179, Grant 61125305, Grant 61170077, Grant 61362031, Grant 61332011, and Grant 61370163; in part by the General Research Fund of Research Grants Council of Hong Kong under Project 531708; 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 JCYJ20120613153352732, Grant JCYJ20120613134843060, and Grant JCYJ20130329152024199.

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

Y. Xu is with the Bio-Computing Research Center and Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: yongxu@ymail.com).

Q. Chan is with the Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: qingcai.chen@hitsz.edu.cn).

J. Yang is with the School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@njust.edu.cn).

D. Zhang is with the Department of Computing, Biometrics Research Centre, Hong Kong Polytechnic University, 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/TNNLS.2013.2297381

form and, simultaneously, introducing the L_1 norm for sparse feature selection is appreciated and may further improve the algorithm's (i.e., MSPCA) performance. Differing from all the multilinear learning methods mentioned above, our multilinear learning method is subjected to the sparseness constraint imposed by the L_1 norm. MSPCA performs sparse dimensionality reduction in all tensor modes and captures most of the variation of the original tensors instead of the vectors.

The main contributions of this brief are as follows. MPCA is first rewritten into multilinear regression forms according to the optimization technique of mode-k flattening of the nth-order tensor. Then, the multilinear regression models are relaxed to an equivalent optimization problem for sparse principal component learning using the L_1 and L_2 norms penalty. Thus, an efficient multilinear sparse principal learning method called MSPCA is proposed for feature extraction. The optimal multilinear sparse principal component vectors are obtained from an iterative algorithm using elastic net regression and singular value decomposition (SVD) on tensor data instead of vectors. Thus, MSPCA is significantly different from the existing multilinear learning methods (such as MPCA and those in [11]–[18] etc.), which use eigen decomposition to compute the optimal multilinear projections.

The rest of this brief is organized as follows. In Section II, MPCA and SPCA are reviewed. In Section III, MSPCA algorithm and related analyses are described. Experiments are carried out to evaluate MSPCA algorithm in Section IV, and the conclusion is given in Section V.

II. Brief Review of SPCA and MPCA

In this section, we give some basic multilinear notations, definitions and operations at first and then briefly review SPCA and MSPCA.

A. Tensor Fundamentals and its Notations

In this brief, if there is no specification, lowercase italic letters, i.e., α, β, k, i, j, denote scalars, bold lowercase letter, i.e., $\mathbf{u}, \mathbf{v}$, denote vectors, uppercased letters, i.e., U, V, B, X, denote matrices and bold uppercase letters $\mathbf{X}, \mathbf{Y}$ denote the tensors.

Assume that the training samples are represented as the nth-order tensor $\{\mathbf{X}_i \in R^{m_1 \times m_2 \times \cdots \times m_n}, i = 1, 2, \ldots, N\}$, where N denotes the total number of the training samples. We need the following notations and definitions similar to [12]–[14], [19], [27].

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

For the intuitive representation of the mode-k flattening of the nth-order tensor, the readers are referred to the figures in [19].

Definition 2: The mode-k product of tensor $\mathbf{X}$ with matrix $U \in R^{m'_k \times m_k}$ is defined as $\mathbf{Y} = \mathbf{X} \times_k U$, where $\mathbf{Y}_{i_1, \ldots, i_{k-1}, i, i_{k+1}, \ldots, i_n} = \sum_{j=1}^{m_k} \mathbf{X}_{i_1, \ldots, i_{k-1}, j, i_{k+1}, \ldots, i_n} U_{i,j}$ $(j = 1, \ldots, m'_k)$.

Without loss of generality, we assume that the training tensor samples are centered and still denoted as $\{\mathbf{X}_i \in R^{m_1 \times m_2 \times \cdots \times m_n}, i = 1, 2, \ldots, N\}$. With the above preparations, we begin to review the related algorithms.

B. Sparse PCA

It is well known that PCA can be formulated as a ridge regression optimization problem. Thus, the sparse projections of SPCA can be obtained by imposing the lasso constraint on the regression problem. That is, SPCA considers the following elastic net regularization problem:

$$(\hat{Q}, \hat{P}) = \arg\min \sum_i \left\| x_i - Q P^T x_i \right\|^2 + \alpha \left\| P \right\|_F^2 + \sum_j^{d_1} \beta_j \left| \mathbf{p}_1^j \right| \quad \text{s.t.} \quad Q^T Q = I_d \tag{1}$$

where x_i is the high-dimensional vector concatenated by the columns/rows of tensor $\mathbf{X}_i$, and Q and P are the $d \times (m_1 \times m_2 \times \cdots \times m_n)$ matrices, and $\alpha, \beta_j \geq 0$ is used for penalizing the loadings of different principal component vectors.

The problem defined in (1) can be solved by an alternative minimization algorithm to compute the optimal $\hat{Q}$ and $\hat{P}$. Then, the sparse principal component vectors are $\hat{\boldsymbol{p}}_i / \left\| \hat{\boldsymbol{p}}_i \right\|$, $(i = 1, 2, \ldots, d)$, which can be used for feature extraction.

C. Multilinear PCA

The purpose of the MPCA is to obtain a set of multilinear transformations (or projections) $V_i \in R^{m_i \times d_i}$ $(d_i \leq m_i, i = 1, 2, \ldots, n)$ that map the original high-order tensor data into a low-order tensor space

$$\mathbf{Y}_i = \mathbf{X}_i \times_1 V_1^T \times_2 V_2^T \cdots \times_n V_n^T. \tag{2}$$

The objective function of the MPCA is to directly maximize the total scatter matrix on the subspace $V_i \in R^{m_i \times d_i}$ $(i \neq k)$ [19]

$$\max_{V_k^T V_k = I_k} \operatorname{tr}\left(V_k^T S_T^k V_k\right) = \max_{V_k^T V_k = I_k} \operatorname{tr}\left[V_k^T \left(\sum_i^N X_i^k X_i^{kT}\right) V_k\right] \tag{3}$$

where $S_T^k = \sum_i^N X_i^k X_i^{kT}$ and X_i^k be the mode-k flattening matrix of $\mathrm{X}_i^k \Leftarrow \mathbf{X}_i \times_1 V_1^T \times \cdots \times_{k-1} V_{k-1}^T \times_{k+1} V_{k+1}^T \times \cdots \times_n V_n^T$. The optimal projections of MPCA can be obtained from the SVD or eigen decomposition

$$S_T^k V_k = V_k D_k \tag{4}$$

where $V_k = [\mathbf{v}_k^1, \mathbf{v}_k^2, \ldots, \mathbf{v}_k^{d_k}]$ is the eigenvector matrix and $D_k = \operatorname{diag}(\lambda_k^1, \lambda_k^2, \ldots, \lambda_k^{d_k})$ is the eigenvalue matrix of S_T^k, where $\lambda_k^1 \geq \lambda_k^1 \geq \ldots \geq \lambda_k^{d_k}$ and λ_k^j is the eigenvalue corresponding to the eigenvector $\mathbf{v}_k^j$. The optimal projection matrix for mode-k is composed of the eigenvectors corresponding to the first d_k largest eigenvalues. That is, $V_k = [\mathbf{v}_k^1, \mathbf{v}_k^2, \ldots, \mathbf{v}_k^{d_k}]$ is the projection matrix of MPCA for each mode.

III. Multilinear SPCA

In this section, the multilinear regression for MPCA is presented at first. Then, it is modified for multilinear sparse principal component learning.

A. Multilinear Ridge Regression for MPCA

It is known that PCA can be represented in the regression form. Similarly, we extend the single linear regression into multilinear regression, which aims to minimize the tensor reconstruction error with the L_2 norm penalty. Let $B_i \in R^{m_i \times d_i} (i = 1, 2, \ldots, N)$ and

$$\begin{aligned} J(B_1, B_2, \ldots, B_n) &= \sum_i \left\| \mathbf{X}_i - \mathbf{X}_i \times_1 B_1 B_1^T \times_2 B_2 B_2^T \ldots \times_n B_n B_n^T \right\|_F^2 \\ &\quad + \sum_j \alpha_j \left\| B_j \right\|_F^2. \end{aligned} \tag{5}$$

MSPCA Algorithm

Input: Tensor samples $\{\mathbf{X}_i \in R^{m_1 \times m_2 \times \cdots \times m_n}, i = 1,2,...,N\}$, the numbers of iterations $T_{\max}$, the dimensions $d_i(\le m_i), i = 1,2,...,n$

Output: Multilinear sparse subspace $U_1, U_2, \cdots, U_n$ ($i = 1,2,...,N$)

Step 1: Center the training input samples.

Step 2: Initialize $U_k^1, B_k^1|_{k=1}^n$ as arbitrary columnly-orthogonal matrices.

Step 3: For $t = 1 : T_{\max}$ do

For $k = 1 : n$ do

*Compute $\mathbf{X}_i^k$: $\mathbf{X}_i^k = \mathbf{X}_i \times_1 U_1^{tT} \cdots \times_{k-1} U_{k-1}^{tT} \times_{k+1} U_{k+1}^{tT} \cdots \times_n U_n^{tT}$

*Perform the mode- k flattening of the n th-order tensors $\mathbf{X}_i^k$ to matrices: $X_i^k \Leftarrow_k \mathbf{X}_i^k$

*Repeat the following two steps until $\widehat{U}_k$ converges

-Solve the Elastic Net problem: $\widehat{U}_k \leftarrow \arg\min \sum_i \left\| X_i^k - B_k^t U_k^T X_i^k \right\|_F^2 + \alpha \left\| U_k \right\|_F^2 + \sum_j^{d_k} \beta_{k,j} \left| \mathbf{u}_k^j \right|$

-Perform SVD of $(\sum_i X_i^k X_i^{k^T})\widehat{U}_k = \overline{U}_k \overline{D}_k \overline{V}_k$, and update $B_k^t \leftarrow \overline{U}_k \overline{V}_k^T$.

*Normalize $\widehat{U}_k$, i.e. let $U_k(:,s) = \widehat{U}_k(:,s) / \left\| \widehat{U}_k(:,s) \right\|, s = 1 : d_k$

*If $t \ge 2$ and $\left| S_T^k(t+1) - S_T^k(t) \right| / S_T < \varepsilon$ then break.

End

End

Step 4: Output the multilinear sparse subspaces $U_1, U_2, \cdots, U_n$ ($i = 1,2,...,N$) .

Then, the multilinear ridge regression optimization problem of MPCA can be stated as follows:

$$\min_{B_1, B_2, \ldots, B_n} J(B_1, B_2, \ldots, B_n) \tag{6}$$
$$\text{s. t.} \quad B_1^T B_1 = I_1 , \ldots , B_n^T B_n = I_n .$$

To the best of our knowledge, there exists no closed-form solution for such complex objective function. Fortunately, the optimization problem can be converted into a problem to independently determine n subspaces B_ns that minimize the construction errors of the mode-k flattening of the nth-order tensors using an iterative algorithm similar to MPCA. Therefore, in the following sections, we only focus on the mode-k flattening of the nth-order tensors. Suppose $B_1, B_2, \ldots, B_{k-1}, B_{k+1}, \ldots, B_n$ are known, from Theorem 1 in [11] the minimization problem in (6) converts to the following optimization problem with the single constraint:

$$\min_{B_k} \sum_i \left\| X_i^k - B_k B_k^T X_i^k \right\|_F^2 + \alpha_k \|B_k\|_F^2 \tag{7}$$
$$\text{s.t.} \quad B_k^T B_k = I_k .$$

From (7), we have

$$\min_{B_k} \sum_i \|X_i^k - B_k B_k^T X_i^k\|_F^2 + \alpha_k \|B_k\|_F^2$$
$$= \min_{B_k} 2\mathrm{tr}(S_T^k) - \mathrm{tr}\big[B_k^T (S_T^k + \alpha_k I_k) B_k\big]. \tag{8}$$

Minimizing (8) is equivalent to

$$\max_{B_k^T B_k = I_k} \mathrm{tr}\big[B_k^T (S_T^k + \alpha_k I_k) B_k\big]. \tag{9}$$

Since $\alpha_k I_k$ does not affect the eigenvectors, (4) and (9) have the same solutions, that is V_k is the optimal solution for (9). Therefore, the eigenvectors of MPCA can be derived from multilinear regression.

B. Model Relaxation for MSPCA

Since the multilinear feature extraction methods focus on the optimization problem of the mode-k flattening of the nth-order tensors, in what follows, we pay attention to the multilinear regression problem (7). Let $U_k \in R^{m_k \times d_k}$, to obtain the sparse regression model for sparse principal component vectors, we relax (7) to a new regression problem

$$\min_{U_k, B_k} \sum_i \|X_i^k - B_k U_k^T X_i^k\|_F^2 + \alpha_k \|U_k\|_F^2 \tag{10}$$
$$\text{s. t.} \quad B_k^T B_k = I_k .$$

The following theorem reveals the close relationship between (7) [or (6)] and (10).

Theorem 1: For any given $\alpha_k > 0$, the optimization problems in (7) and (10) have the same solution for variable B_k. Let $(\hat{B}_k, \hat{U}_k)$ be the optimal solutions of (10), where $\hat{U}_k = [\mathbf{u}_k^1, \mathbf{u}_k^2, \ldots, \mathbf{u}_k^{d_k}]$, then $\mathbf{u}_k^j \propto \mathbf{v}_k^j$ $(1 \le j \le d_k)$.

Proof: The proof is shown in the Supplementary Appendix.

Theorem 1 reveals the close relationship between the multilinear regression problem and the objective function of MPCA for any given mode k. Since the optimal solutions of (6) are given by (7) and (10) also has the same solution as (6). This relaxation provides us a tractable method for using the L_1 norm penalty to compute the sparse vectors. Therefore, to obtain the multilinear sparse principal vectors, the lasso penalty is imposed on the regression representation of mode-k flattening of the MPCA regression criterion. Finally, MSPCA criterion is defined as follows:

$$\min_{B_k, U_k} \sum_i \|X_i^k - B_k U_k^T X_i^k\|_F^2 + \alpha_k \|U_k\|_F^2 + \sum_j^{d_k} \beta_{k,j} |\mathbf{u}_k^j| (\forall k) \tag{11}$$
$$\text{s.t.} \quad B_k^T B_k = I_k$$

where $\beta_{k,j} \ge 0$ is used for penalizing the loadings of different principal component vectors. Obviously, if all of the $\beta_{k,j}$s are equal to zero, we have the exact MPCA solutions. In MSPCA algorithm, we usually suppose $\beta_{k,j} > 0$, which results in the multilinear sparse principal vectors.

C. Solutions of MSPCA

This section presents how to compute the sparse solutions of MSPCA. According to (11), we have

$$\begin{aligned}&\sum_i \|X_i^k - B_k U_k^T X_i^k\|_F^2 + \alpha \|U_k\|_F^2 + \sum_j^{d_k} \beta_{k,j} |\mathbf{u}_k^j| \\ &= \mathrm{tr}\left(\sum_i X_i^k X_i^{k^T}\right) + \sum_j \left(\mathbf{u}_k^{jT}\left[\left(\sum_i X_i^k X_i^{k^T}\right) + \alpha\right]\mathbf{u}_k^j \right.\\ &\quad \left. -2b_k^{jT}\left(\sum_i X_i^k X_i^{k^T}\right)\mathbf{u}_k^j + \beta_{k,j}|\mathbf{u}_k^j|\right) \\ &= \mathrm{tr}(S_T^k) + \sum_j \left(\mathbf{u}_k^{\mathrm{jT}}(S_T^k + \alpha)\mathbf{u}_k^j - 2\mathbf{b}_k^{jT} S_T^k \mathbf{u}_k^j + \beta_{k,j}|\mathbf{u}_k^j|\right).\end{aligned} \tag{12}$$

It can be observed that if B_k is given then the optimal sparse solutions of (12) are exactly those of the following m_k' independent naive elastic net problems:

$$\mathbf{u}_k^{jT}(S_T^k + \alpha)\mathbf{u}_k^j - 2\mathbf{b}_k^{jT} S_T^k \mathbf{u}_k^j + \beta_{k,j}|\mathbf{u}_k^j| \quad j = 1, \ldots, m_k'. \tag{13}$$

On the other hand, we also have

$$\begin{aligned}&\sum_i \|X_i^k - B_k U_k^T X_i^k\|_F^2 + \alpha \|U_k\|_F^2 + \sum_j^{d_k} \beta_{k,j} |\mathbf{u}_k^j| \\ &= \mathrm{tr}\left(\sum_i X_i^k X_i^{k^T}\right) - 2\mathrm{tr}\left[B_k^T\left(\sum_i X_i^k X_i^{k^T}\right)U_k\right] \\ &\quad + \mathrm{tr}\left[U_k^T\left(\sum_i X_i^k X_i^{k^T} + \alpha I_k\right)U_k\right] + \sum_i \beta_{k,j}|\mathbf{u}_k^j|\end{aligned} \tag{14}$$

when U_k is known and fixed, the 1-, 3-, and 4-th terms of (14) are constants and thus can be ignored. Finally, the problem of minimizing (14) becomes the following maximization problem:

$$\max\ \mathrm{tr}\left[B_k^T\left(\sum_i X_i^k X_i^{k^T}\right)U_k\right] \quad \text{s. t.} \quad B_k^T B_k = I_k. \tag{15}$$

According to Theorem 4 in [26], the optimal solution of the above maximum problem for the given U_k is

$$B_k^* = \bar{U}\bar{V}^T \tag{16}$$

where $\bar{U}$and $\bar{V}$ are the SVD decomposition of $(\sum_i X_i^k X_i^{k^T})U_k$, i.e., $(\sum_i X_i^k X_i^{k^T})U_k = \bar{U}\bar{D}\bar{V}$. The details of MSPCA algorithm are given in Supplementary Appendix.

D. Computational Complexity and Convergence of MSPCA

1) 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 with the feature dimension m^n. The complexity of PCA is $O(m^{3n})$. The complexity of SPCA is $O(Tm^{3n} + Nm^{2n})$, where T is the number of iterations in elastic net. The total complexity of MPCA (or CSA) is $tO((n+1)Nnm^{n+1} + nm^3)$, where t denotes the number of iterations. The complexity of MSPCA is $tO(n^2Nm^{n+1} + nNm^{n+1} + nTm^3)$ at most. Although many loops are required for MSPCA in its optimization, MSPCA is still more efficient than PCA and SPCA owing to the facts: 1) the inner loops of elastic net are convergent very fast since it operates on the very low-dimensional vectors (or small size matrices) and 2) as shown in Fig. 1 (e) and (f), the outer loops are also convergent within several iterations. Obviously, the complexity of MSPCA will be slightly bigger than that of MPCA even with the same t. However, computing the sparse principal component vectors is only needed in the training phase in pattern recognition tasks, therefore it can be done offline and the additional computational cost is not considered a distinct disadvantage of the proposed method in this case.

2) Convergence of MSPCA: Since (10) gives the same solution space as (7) and (11) indicates that U_k is the optimal sparse approximation to B_k, MSPCA is convergent fast. Due to the existence of the L_1 norm penalty term in MSPCA, it is difficult to strictly prove the convergence of MSPCA algorithm. However, experimental results shown in Fig. 1 indicate that MSPCA converges very fast.

3) Termination Criterion of MSPCA: MSPCA is the sparse version of MPCA. Thus, the scatter characterized by the multilinear sparse principal vectors is a good convergence criterion of MSPCA, which will be shown in the experimental section (i.e., Section IV). Let S_T be the total sample scatter value, $S_T^k(t)$ and $S_T^k(t+1)$ be the total scatter value characterized by the projection matrices in t- and $t+1$-th iterations, respectively. If $\left|S_T^k(t+1) - S_T^k(t)\right| / S_T < \varepsilon$, where ε is a small constant such as $\varepsilon = 0.001$, then MSPCA algorithm can be regarded as convergent and the iteration procedure is terminated. Experimental results presented in Section IV show that the proposed MSPCA converges very fast for tensor data. Fig.1 (e) and (f) shows this property.

IV. Experiment

In this section, a set of experiments are presented to evaluate the proposed MSPCA algorithm for image recognition tasks when the order of the data n is equal to 2. The Yale face database (http://www.cvc.yale.edu/projects/ yalefaces/yalefaces.html) was used to explore the properties of MSPCA, including the convergence, robustness on the variations in expression and illumination. The Face Recognition Technology (FERET) face database [29] was used to evaluate the performance of these methods when face poses and lighting conditions varied. The COIL-20 image database (http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php) was used to evaluate the robust performance of these methods when the objective's images varied in rotations. Finally, the performance of higher order case ($n = 3$) of MSPCA was test on Weizmann action database [30]. The details about the databases used in this brief are described in the Supplementary Appendix. The nearest neighborhood classifier with Euclidean distance was used in all the experiments. The related Matlab codes of this brief can be available from Yong Xu's homepage (http://www.yongxu.org/lunwen.html).

A. Exploration for the Properties of MSPCA

First, we explore some properties of MSPCA, including the convergence property, the variations of recognition rate versus the number of iterations, and the relationship between the sparsity (i.e., the number of nonzero loadings), the number of the projections and the recognition rates. Fig. 1 shows these properties of MSPCA on the Yale face database by randomly selecting four images per individual for training and the remaining for test.

From Fig. 1(a), it can be found that MSPCA can achieve its best performance using only very small number of nonzero elements (usually within $2 \sim 8$ nonzero elements) on its projection. Fig. 1(b) shows that MSPCA can use less than 10 projection vectors in each U_k ($k = 1, 2$) to achieve its top recognition rate, which indicates the dimensionality reduction ability in face recognition. Fig. 1(c) shows the recognition rate versus the variation of the key parameter alpha in MSPCA, which indicates that if the L_2 norm penalty parameter is set to be 0 (i.e., without L_2 norm penalty term), the recognition rate is relatively low. This indicates that combining the L_1 and L_2 norms penalty in MSPCA can enhance the performance. As it is shown in [24] and [26], LASSO (only using the

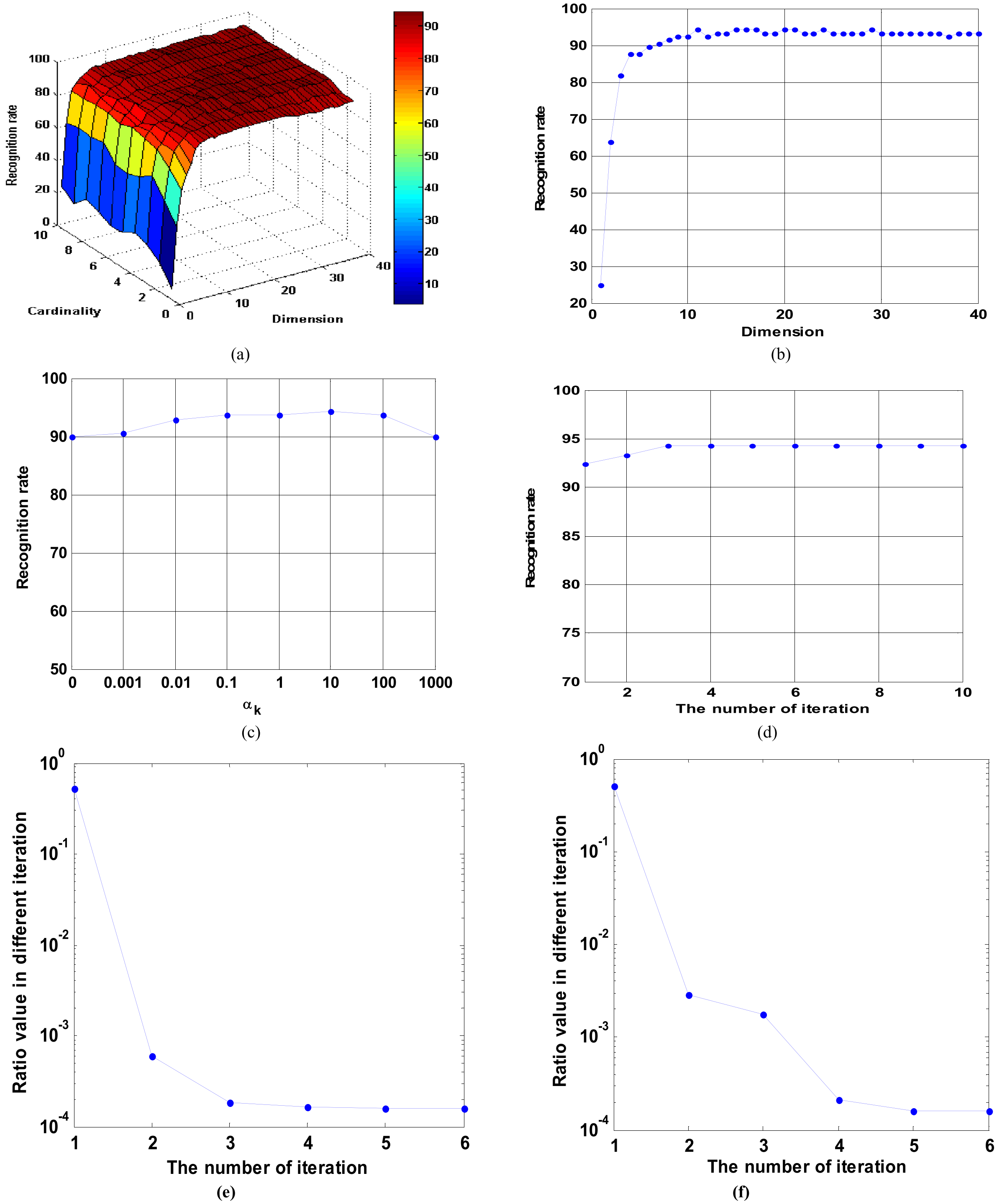

Fig. 1. Some properties of MSPCA. (a) Recognition rate versus the variations of dimension and the number of cardinality. (b) Recognition rate versus the variation of dimension. (c) Recognition rate versus the variation of parameter alpha. (d) Variation of recognition rate versus the number of iteration. (e) Convergence of MSPCA on Yale face database. (f) Convergence on COIL20 image database.

L_1 norm in regression) has some uncertainty in feature selection when there are several factors to be selected (LASSO randomly selects one variable in this case), leading to a certain degree of uncertainty in the projections. Thus, the performance will be degraded. However, elastic net overcomes this drawback in LASSO and leads to obtain more informative principal vectors. Fig. 1(d) shows the recognition rate versus the variation of the number of iteration, which indicates that the recognition rate is not affected when the algorithm achieves three iterations. Fig. 1(e) and (f) shows the ratio value (i.e., $|S_T^k(t+1) - S_T^k(t)|/S_T$) of the variations of the scatter in different number of iterations. Usually, MSPCA will converge within several iterations, which is similar to MPCA. Similar properties can also be found in the other databases, but we will not present them for saving space.

TABLE I
PERFORMANCE (AVERAGE RECOGNITION RATE, DIMENSION, AND STANDARD DEVIATION) OF THE COMPARED METHODS ON THE YALE FACE DATABASE

Training samples	PCA	PCA +LDA	SPCA	SPCA +LDA	2DPCA	2DPCA +LDA	MPCA	MPCA +LDA	MSPCA	MSPCA +LDA
4	85.80 19 ±2.45	90.90 14 ±3.07	86.81 39 ±2.39	90.14 14 ±2.99	88.00 40×14 ±2.70	90.33 14 ±4.02	87.81 16×16 ±2.48	91.71 14 ±3.17	93.12 21×21 ±1.35	93.92 14 ±1.63
5	86.22 28 ±4.56	91.33 14 ±4.88	87.21 32 ±4.38	91.56 14 ±4.04	88.78 40×14 ±4.18	91.55 14 ±2.95	88.89 14×14 ±4.71	92.33 14 ±3.36	94.78 15×15 ±1.39	95.81 14 ±1.94

TABLE II
PERFORMANCE (AVERAGE RECOGNITION RATE, DIMENSION, AND STANDARD DEVIATION) OF THE COMPARED METHODS ON THE FERET FACE DATABASE

Training samples	PCA	PCA +LDA	SPCA	SPCA +LDA	2DPCA	2DPCA +LDA	MPCA	MPCA +LDA	MSPCA	MSPCA +LDA
4	59.90 180 ±10.62	68.51 199 ±11.93	60.31 195 ±10.78	69.93 199 ±9.38	60.47 40×31 ±3.29	68.61 199 ±5.97	64.36 15×15 ±10.72	73.63 199 ±4.91	71.21 17×17 ±5.06	78.95 199 ±4.95
5	62.23 195 ±12.11	73.7 199 ±8.00	62.75 200 ±12.94	75.20 199 ±6.45	62.95 40×25 ±12.12	73.70 199 ±8.00	73.91 16×16 ±4.91	81.52 199 ±5.53	80.72 18×18 ±4.50	86.98 199 ±3.50

B. Experimental Settings

PCA, SPCA, 2-DPCA, MPCA, and MSPCA are used for feature extraction to compare their performance. PCA and SPCA are operated on the high-dimensional vectors of the images, 2-DPCA, MPCA, and MSPCA are directly operated on the image matrices. For face feature extraction, MPCA and MSPCA were used as the second-order (matrix-based) forms, and the third-order of MSPCA will be explored in action recognition.

In each experiment presented in the following sections, the databases are divided into training, validation, and test sets. In each experiment, four and five (five and six on Weizmann action database) samples of each individual/class are randomly selected as the training set, and one half of the remaining is randomly selected as validation set and the other half as test set. For each run, the optimal parameters determined by the validation set (i.e., the parameters corresponding to the best recognition rate on the validation set) were used to learn the projections for feature extraction and classification. The experiments were repeated 10 times and the average recognition rates, the standard deviations and the optimal dimensionalities one the test set are reported. For simplicity, in the experiments, we set $\alpha_1 = \alpha_2 = \alpha_3 = \bar{\alpha}$ and the optimal coefficients of L_2 norm penalty term $\bar{\alpha}$ was selected from 0.01, 0.1, …, 1000. The coefficients of L_1 norm term (i.e., β or β_k) are automatically determined by the elastic net algorithm since the elastic net algorithm can compute the optimal solution path for any given $\bar{\alpha}$ (or α_k) (thus, it is not necessary to manually set the parameter β_k). The cardinality of each sparse projection was selected from 1 to m_i using the validation set. The optimal dimensions of the subspace for 2-DPCA, MPCA, and MSPCA were selected from $1 \sim m_k$. The optimal dimensions of the subspaces for PCA and SPCA were selected from 1 to $N - 1$.

Since these methods mentioned above are all unsupervised, in the experiments, we also report the performance using classical linear discriminant analysis (LDA) for further supervised feature extraction (this two stage strategy is denoted as *+LDA and * denotes the above principal component learning methods). In this case, the samples are projected to the $C - 1$ dimensions subspace, where C denotes the number of classes.

C. Performance of MSPCA

In evaluation of the performance of the proposed algorithm in feature extraction and recognition, the average recognition rates, standard deviations and the optimal dimensionalities achieving the top accuracies are reported in Tables I–IV. Note that 2-DPCA cannot be used in action recognition, we only report the experimental results of the remaining methods in Table IV. PCA and SPCA are operated on the very high-dimensional vectors ($32 \times 24 \times 10 = 7680$), which are obtained by concatenating column by column of the tensor data. MPCA and MSPCA are directly operated on the 3rd tensors.

To test the computation times of PCA, SPCA, 2-DPCA, MPCA, and MSPCA, all the experiments are run on a personal computer [Intel (R) Core i7, CPU: 2.67 GHz, RAM: 4 GB] using MATLAB. We take COIL-20 database as an example (using four images per individual in the training stage) to compare the efficiency of these methods. The computational times (in seconds) of PCA, 2-DPCA, SPCA, MPCA, and MSPCA are 2.96, 0.12, 63.45, 0.23, and 0.35 s, respectively. It can be found that when SPCA is extended to MSPCA, the efficiency is greatly improved in learning the sparse principal component vectors.

D. Observations and Discussions

Though there are variations in illumination and facial expression, from the recognition rates, it can be found that MSPCA outperforms the other principal component analysis algorithms. SPCA usually performs better than PCA, and MSPCA also performs better than MPCA. This shows that MSPCA enhances its robustness by introducing the L_1 and L_2 norms penalty regression to select the most important factors/variables for feature extraction. Another reason why MSPCA and MPCA perform better than SPCA and PCA, respectively, is that the structure information embedded in the higher order tensors does enhance the feature extraction abilities of one method. Thus, the higher order tensor extension of a feature extraction method is a tractable way to enhance the performance for feature extraction.

For the two stage methods for feature extraction and recognition, LDA usually significantly enhance the performance on face recognition. However, when there are large rotations in the objective images,

TABLE III

PERFORMANCE (AVERAGE RECOGNITION RATE, DIMENSION, AND STANDARD DEVIATION) OF THE COMPARED METHODS ON THE COIL-20 FACE DATABASE

Training samples	PCA	PCA +LDA	SPCA	SPCA +LDA	2DPCA	2DPCA +LDA	MPCA	MPCA +LDA	MSPCA	MSPCA +LDA
4	81.22 17 ±3.35	80.83 19 ±3.22	82.26 17 ±3.34	82.18 19 ±3.32	82.22 32×8 ±3.15	82.69 19 ±2.81	83.11 6×6 ±3.16	83.17 19 ±2.56	86.15 6×6 ±3.38	86.33 19 ±2.27
5	83.49 19 ±2.90	83.68 19 ±2.77	84.62 20 ±2.96	85.00 19 ±3.20	85.49 32×9 ±3.07	85.11 19 ±2.68	85.16 6×6 ±3.20	85.39 19 ±3.39	88.38 6×6 ±2.91	88.39 19 ±2.96

TABLE IV

PERFORMANCE (AVERAGE RECOGNITION RATE, DIMENSION, AND STANDARD DEVIATION) OF THE COMPARED METHODS ON THE WEIZMANN ACTION DATABASE

Training samples	PCA	PCA +LDA	SPCA	SPCA +LDA	MPCA	MPCA +LDA	MSPCA	MSPCA +LDA
5	72.57 22 ±3.07	75.11 9 ±4.07	73.34 40 ±3.74	76.09 9 ±4.17	70.94 9^3 ±4.50	71.68 9 ±2.63	77.69 7^3 ±3.92	78.63 9 ±3.17
6	76.09 17 ±2.85	77.71 9 ±2.69	76.35 17 ±3.35	78.04 9 ±3.16	71.27 9^3 ±3.72	73.53 9 ±3.73	80.00 9^3 ±3.28	80.48 9 ±2.38

LDA can hardly enhance the accuracy due to the large within-class scatter value.

When MPCA performs less effectively for some applications such as action recognition, MSPCA still achieves the best performance among the compared methods. This shows that MSPCA is more stable and has stronger generalization ability and robustness in feature extraction than MPCA and other principal component learning methods.

V. CONCLUSION

This brief extends the SPCA or MPCA to multilinear sparse case named MSPCA. We first convert the tensor-based MPCA criterion into the multilinear ridge regression, which is then relaxed for the multilinear sparse principal component learning by imposing the L_1 norm penalty. The optimal sparse vectors of MSPCA can be efficiently computed by iterative procedures using the sparse regression method. Thus, this brief addresses the problem of multilinear sparse principal component learning with the L_1 and L_2 norms regression for arbitrary high-order tensors. Computational and convergent analyses were presented to show the properties of the MSPCA. The experiments on second- and third-order tensor databases indicate that MSPCA performs better than PCA, SPCA, 2-DPCA, and MPCA in feature extraction.

APPENDIX A
PROOF OF THEOREM 1

Proof: From (10), we have

$$\begin{aligned}\sum_i \|X_i^k - B_k U_k^T X_i^k\|_F^2 + \alpha_k \|U_k\|_F^2 &= \mathrm{tr}\left(\sum_i X_i^k X_i^{k^T}\right) - 2\mathrm{tr}\left[B_k^T\left(\sum_i X_i^k X_i^{k^T}\right)U_k\right] \\ &\quad + \mathrm{tr}\left[U_k^T\left(\sum_i X_i^k X_i^{k^T} + \alpha I_k\right)U_k\right] \\ &= \mathrm{tr}(S_T^k) - 2\mathrm{tr}(B_k^T S_T^k U_k) + \mathrm{tr}[U_k^T(S_T^k + \alpha I_k)U_k]. \end{aligned} \tag{A1}$$

For the fixed B_k, using Lagrange multiplier method, we know that the above quantity is minimized at

$$\hat{U}_k = \left(S_T^k + \alpha I_k\right)^{-1} S_T^k B_k. \tag{A2}$$

Substituting (A2) into (A1), we have

$$\sum_i \|X_i^k - B_k U_k^T X_i^k\|_F^2 + \alpha_k \|U_k\|_F^2 = \mathrm{tr}(S_T^k) - \mathrm{tr}(B_k^T S_T^k (S_T^k + \alpha I_k)^{-1} S_T^k B_k). \tag{A3}$$

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

$$\arg\max_{B_k} \mathrm{tr}(B_k^T S_T^k (S_T^k + \alpha I_k)^{-1} S_T^k B_k) \quad \text{s.t.} \quad B_k^T B_k = I_k. \tag{A4}$$

Since V_k is the solution to (9), and therefore it is also the solution to (7). Denoting the SVD of $S_T^k = \widehat{V}_k \widehat{D}_k \widehat{V}_k^T$, we have

$$S_T^k (S_T^k + \alpha I_k)^{-1} S_T^k = \widehat{V}_k \frac{\widehat{D}_k^2}{\widehat{D}_k + \alpha I_k} \widehat{V}_k^T. \tag{A5}$$

Form (A5), It is straightforward to find that $B_k = V_k = [\mathbf{v}_k^1, \mathbf{v}_k^2, \ldots, \mathbf{v}_k^{d_k}]$ is the solution of (A4). This shows (7) and (10) have the same solution. Therefore, from (A2), we obtain

$$\hat{U}_k = (S_T^k + \alpha I_k)^{-1} S_T^k B_k = V_k \frac{\widehat{D}_k(1:d_k, 1:d_k)}{\widehat{D}_k(1:d_k, 1:d_k) + \alpha I_{d_k}} \tag{A6}$$

where $\widehat{D}_k(1 : d_k, 1 : d_k)$ denotes the submatrix of $\widehat{D}_k$. Thus, $\mathbf{u}_k^j \propto \mathbf{v}_k^j$ $(1 \le j \le d_k)$.

APPENDIX B
DESCRIPTION OF THE DATA SETS

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

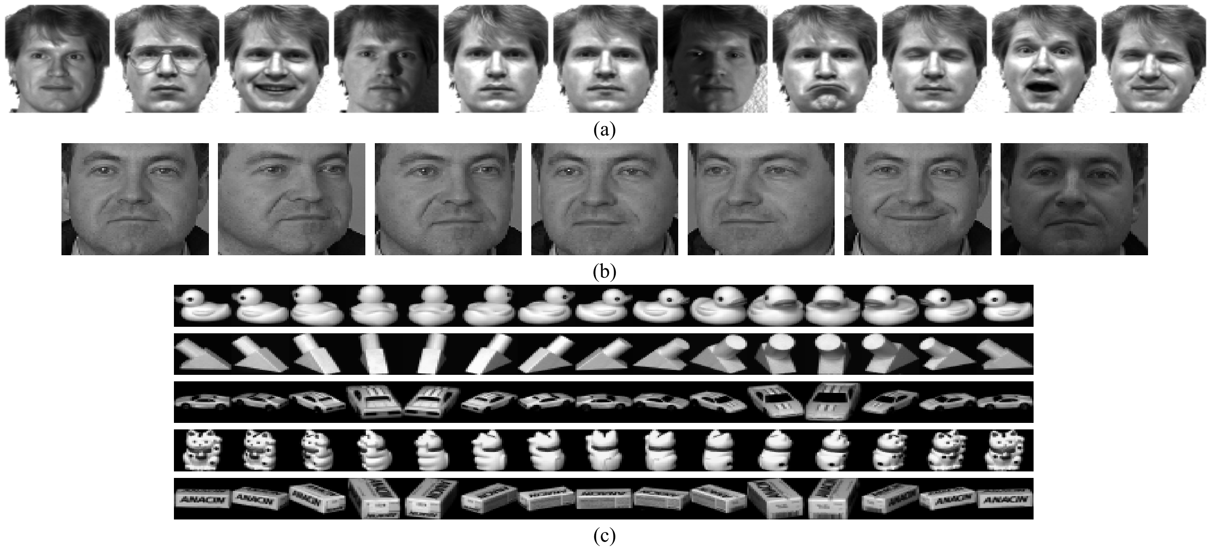

Fig. A1. Image samples used in the experiments. (a) Yale face database. (b) FERET face database. (c) COIL-20 object image database.

was manually cropped and resized to 50 × 40 pixels. Fig. A1(a) shows the sample images of one person in the Yale 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 [29]. 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 1400 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 images were preprocessing using histogram equalization at first and then the pixel values were normalized within the area of [0, 1]. The sample images of one person are shown in Fig. A1(b).

The COIL-20 database (http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php) consists of 20 × 72 = 1440 images of 20 objects where the images of each object were taken at pose intervals of 5° (i.e., 72 poses per object). The original images were normalized to 128 × 128 pixels. Each image was converted to a gray scale image of 32 × 32 pixel for computational efficiency in the experiments. Some sample images of four objects are shown in Fig. A1(c).

The experiment was performed on the Weizmann database [30], 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. A2(a).

To represent the spatiotemporal feature of the samples, 10 successive frames of each action were used to extract the temporal feature. Fig. A2(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 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

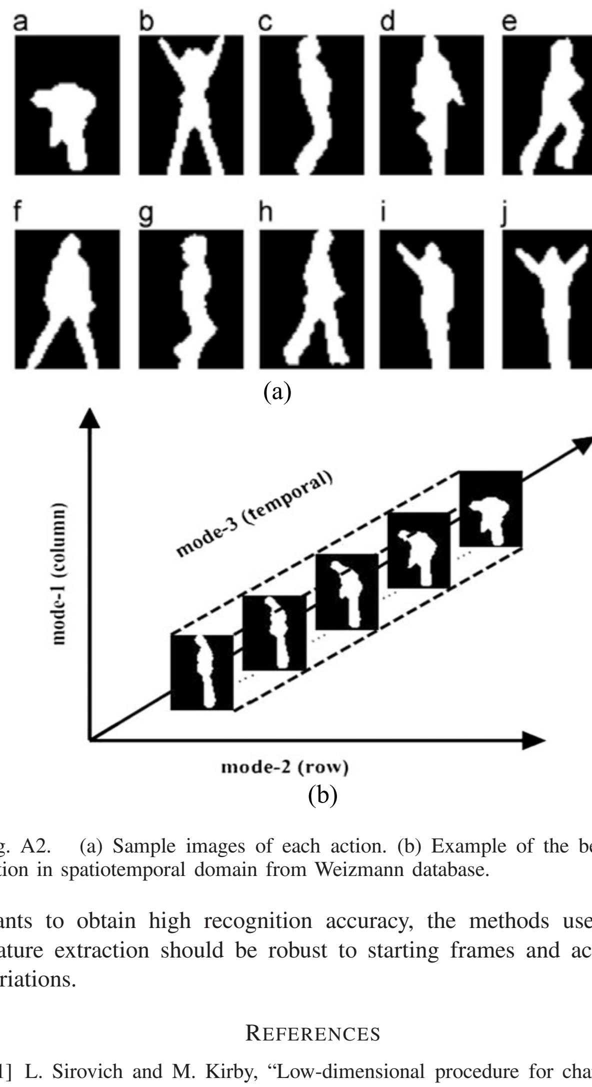

Fig. A2. (a) Sample images of each action. (b) Example of the bending action in spatiotemporal domain from Weizmann database.

wants to obtain high recognition accuracy, the methods used for feature extraction should be robust to starting frames and actions' variations.

REFERENCES

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

[2] 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.

[3] M. Turk, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[4] 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.

[5] 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.

[6] 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, Aug. 2006.

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

[8] D. Zhang and Z. Zhou, "2D2PCA: 2-directional 2-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol. 69, nos. 1–3, pp. 224–231, 2005.

[9] 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.

[10] D. Xu, S. Yan, L. Zhang, H. Zhang, Z. Liu, and H. Shum, "Concurrent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 203–208.

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

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

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

[14] L. Lathauwer, B. De 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. 1, pp. 1324–1342, 2000.

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

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

[17] 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.

[18] 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.

[19] 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.

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

[21] Q. Zhao, G. Zhou, T. Adali, L. Zhang, and A. Cichocki, "Kernelization of tensor-based models for multiway data analysis," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 137–148, Jul. 2013.

[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc., Series B (Statist. Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

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

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

[25] 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.

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

[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] J. Wang, A. Barreto, L. Wang, Y. Chen, N. Rishe, J. Andrian, *et al.*, "Multilinear principal component analysis for face recognition with fewer features," *Neurocomputing*, vol. 73, nos. 10–12, pp. 1550–1555, Jun. 2010.

[29] 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.

[30] Y. Li, C. Fermuller, Y. Aloimonos, and J. Hui, "Learning shift-invariant sparse representation of actions," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 2630–2637.