STPCA: Sparse Tensor Principal Component Analysis for Feature Extraction

Su-Jing Wang[1,2], Ming-Fang Sun[1], Yu-Hsin Chen[2], Er-Ping Pang[1] and Chun-Guang Zhou[1]*
[1]*College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[2]*Institute of Psychpology, Chinese Academy of Sciences, Beijing, 100101, China*
{wangsj08,sunmf09,pangep10}@mails.jlu.edu.cn; yhandrewc@gmail.com; cgzhou@jlu.edu.cn

Abstract

Due to the fact that many objects in the real world can be naturally represented as tensors, tensor subspace analysis has become a hot research area in pattern recognition and computer vision. However, existing tensor subspace analysis methods cannot provide an intuitionistic nor semantic interpretation for the projection matrices. In this paper, we propose Sparse Tensor Principal Component Analysis (STPCA), which transforms the eigen-decomposition problem to a series of regression problems. Since its projection matrices are sparse, STPCA can also address the occlusion problem. Experiment on Georgia tech database and AR database showed that the proposed method outperforms the Multilinear Principal Component Analysis (MPCA) in terms of accuracy and robustness.

1. Introduction

Principal Component Analysis (PCA) is a popular vector subspace analysis method for feature extraction. PCA aims to maximize the variances in the projected subspace by maximizing the trace of covariance matrix. A potential shortage of PCA is that it vectorize a facial image of size m by n to a $(m\times n)$-dimensional vector. In practice, when PCA is applied on the 2D images, one intrinsic problems have been found such as, singularity of within-class scatter matrices, limited available projection directions, high computational cost and a loss of the underlying spatial structure information of the images. In order to address these problems, Lu *et al.* [2] introduces a multilinear principal component analysis (MPCA) for tensor object feature extraction by extended PCA from vector to tensor.

One the common disadvantage amongst all the methods mentioned above is that it is hard to give a physical or semantic interpretation for the projection matrices. However, interpretable models can be obtained via variable selection in multiple linear regression. Thus in recent years, sparse subspace learning has become a hot topic. In [8], Sparse PCA (SPCA) was proposed by applying the least angle regression and elastic net of ℓ_1-penalized regression on regular principal components. However, it is difficult that SPCA is applied on 2D gray images due to the high dimensional vector created through the vectorization. So Xiao and Wang [6] proposed 2D-SPCA, which is directly calculated on image convariance matrix without vectorization. Wang *et al.* used discriminant tensor [5] and sparse discriminant tensor [4] to model color space for face recognition.

In this paper, drawing upon the insights from these methods we propose a Sparse Tensor Principal Component Analysis (STPCA) for feature extraction. The main advantages of STPCA includes:

- STPCA transforms the eigen-decomposition problem into a series of regression problems and can give a intuitionistic or semantic interpretation.
- STPCA has the capability to address the occlusion problem effectively.

2. Sparse Tensor Principal Component Analysis

In this section, we introduce Sparse Tensor Principal Component Analysis to extract the feature of tensor objects. Due to page limit, the concepts and notations of tensor are skipped. For details, please refer to [1]. There are M N-order tensor $\mathcal{X}_m \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_N}$, $m=1,2,\ldots,M$. The STPCA algorithm seeks N sparse projection matrices $\{\mathbf{U}_n \in \mathbb{R}^{I_n \times P_n}, n = 1, \ldots, N\}$ for transformation:

$$\mathcal{Y}_m = \mathcal{X}_m \times_1 \mathbf{U}_1{}^T \times_2 \mathbf{U}_2{}^T \ldots \times_N \mathbf{U}_N{}^T. \qquad (1)$$

which ensures that the projected tensors $\mathcal{Y}_m$ are distributed as far as possible and $\mathbf{U}_n$ is sparse enough. Here, ‘sparsity’ means that $\mathbf{U}_n$ either have a small number of nonzero elements or it has lots of zero elements.

The mean tensor and total scatter are defined by:

$$\overline{\mathcal{X}} = \frac{1}{M}\sum_{m=1}^{M} \mathcal{X}_m \qquad \text{and} \qquad \Psi_{\mathcal{X}} = \sum_{m=1}^{M} \| \mathcal{X}_m - \overline{\mathcal{X}} \|_F^2$$

It is reasonable to maximize the total scatter of projected tensor $\Psi_{\mathcal{Y}}$ as:

$$\{\mathbf{U}_n, n = 1, \ldots, N\} = arg \max_{\mathbf{U}_1, \ldots, \mathbf{U}_N} \Psi_{\mathcal{Y}}. \quad (2)$$

N matrices $\mathbf{U}_n$ need to be simultaneously updated to satisfy the optimal solution of the criterion function. We define *n*-mode scatter matrix $\Phi^{(n)}$ as:

$$\Phi^{(n)} = \sum_{m=1}^{M} (\mathbf{X}_{m(n)} - \overline{\mathbf{X}}_{(n)}) ?\mathbf{U}_n \, \mathbf{U}_n^T \, (\mathbf{X}_{m(n)} - \overline{\mathbf{X}}_{(n)})^T$$

where $\mathbf{U}_n = \mathbf{U}_{n+1} \otimes \cdots \otimes \mathbf{U}_N \otimes \mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_{n-1}$.

Let $\mathbf{U}_n, n = 1, \ldots, N$, be the solution to Eq.(2). Given all the other projection matrices $\mathbf{U}_1, \ldots, \mathbf{U}_{n-1}$, $\mathbf{U}_{n+1}, \ldots, \mathbf{U}_N$, then the matrix $\mathbf{U}_n$ consists of P_n eigenvectors corresponding to the largest P_n eigenvalues of matrix $\Phi^{(n)}$, which satisfies:

$$\Phi^{(n)} \mathbf{u}_p = \lambda \mathbf{u}_p \qquad (3)$$

where $\mathbf{U}_n = [\mathbf{u}_1, \ldots, \mathbf{u}_{P_n}]$. Since $\Phi^{(n)}$ is dependent on $\mathbf{U}_1, \ldots, \mathbf{U}_{n-1}, \mathbf{U}_{n+1}, \ldots, \mathbf{U}_N$, an iterative procedure can be constructed to maximize Eq.(2). For details, please refer to [2].

The aim of STPCA is not only to maximize Eq.(2) but also to make the projection matrices $\mathbf{U}_n$ ($n = 1, \ldots, N$) sparse. So, the criterion function of STPCA is defined as:

$$\begin{aligned} &\{\mathbf{U}_n, n = 1, \ldots, N\} = arg \max_{\mathbf{U}_1, \ldots, \mathbf{U}_N} \Psi_{\mathcal{Y}} \\ &\text{subject to} \quad Card(\mathbf{U}_n) < K_n, \quad n = 1, \ldots, N \end{aligned} \qquad (4)$$

where $Card(\mathbf{U}_n)$ denotes the number of non-zero elements in each column of sparse projection matrix $\mathbf{U}_n$. The only difference between Eq.(2) and Eq.(4) is a sparseness constraint imposed in Eq.(2). The solution to Eq.(4) can be obtained by seeking P_n vectors $\mathbf{b}_p$, such that $\mathbf{b}_p \propto \mathbf{u}_p$, where $\mathbf{u}_p$ is eigenvector in Eq.(3).

We combine all objects $\mathcal{X}_m \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ into a (N +1)-order tensor $\mathcal{H} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N \times M}$.

$$\mathcal{H} = \left\{ \begin{array}{c} (\mathcal{X}_1 - \overline{\mathcal{X}}) \\ (\mathcal{X}_2 - \overline{\mathcal{X}}) \\ \vdots \\ (\mathcal{X}_M - \overline{\mathcal{X}}) \end{array} \right\}$$

Theorem 1 Given $N-1$ projection matrices $\mathbf{U}_1$, $\ldots$, $\mathbf{U}_{n-1}$, $\mathbf{U}_{n+1}$, $\ldots$, $\mathbf{U}_N$, let

$$\mathcal{G} = \mathcal{H} \times_1 \mathbf{U}_1^T \ldots \times_{n-1} \mathbf{U}_{n-1}^T \times_{n+1} \mathbf{U}_{n+1}^T \ldots \times_N \mathbf{U}_N^T$$

Then, $\Phi^{(n)} = \mathbf{G}_{(n)} \mathbf{G}_{(n)}^T$

Proof: The proof is skipped due to the limit pages.

Theorem 2 Let $u_1, u_2, \ldots, u_{P_n}$ denote the eigenvectors of problem Eq.(3) corresponding to the P_n largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{P_n}$ of matrix $\Phi^{(n)}$. Let $\mathbf{A}_{I_n \times P_n} = [\mathbf{a}_1, \ldots, \mathbf{a}_{P_n}]$ and $\mathbf{B}_{I_n \times P_n} = [\mathbf{b}_1, \ldots, \mathbf{b}_{P_n}]$. For any $\lambda > 0$, then $\mathbf{A}$ and $\mathbf{B}$ are the solutions of the following problem:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{\tilde{I}_n} \| \mathbf{H}_{(n)}(:, i) - \mathbf{A}\mathbf{B}^T \mathbf{H}_{(n)}(:, i) \|^2 + \lambda \sum_{j=1}^{P_n} \| \mathbf{b}_p \|^2 \\ &\text{subject to} \quad \mathbf{A}^T \mathbf{A} = \mathbf{I} \end{aligned} \qquad (5)$$

where $\mathbf{H}_{(n)}(:, i)$ denotes the i st column of the mode n unfolding matrix $\mathbf{H}_{(n)}$ and $\tilde{I}_n = I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N \times M$. Then $\mathbf{b}_p \propto \mathbf{u}_p$ for $p = 1, 2, \ldots, P_n$.

Proof: The proof is similar to Theorem 3 in [8]

According to Theorem 2, the generalized eigenvalue of Eq.(3) is transformed to the regression problem of Eq.(5). The regression problem (5) can be solved by iteratively fixing $\mathbf{A}$ and $\mathbf{B}$.

Given a fixed $\mathbf{B}$, we can ignore $\lambda \sum_{p=1}^{P_n} \| \mathbf{b}_p \|^2$ in Eq.(5) and only try to minimize

$$\sum_{i=1}^{\tilde{I}_n} \| \mathbf{H}_{(n)}(:, i) - \mathbf{A}\mathbf{B}^T \mathbf{H}_{(n)}(:, i) \|^2 = \| \mathbf{H}_{(n)}^T - \mathbf{H}_{(n)}^T \mathbf{B}\mathbf{A}^T \|^2$$

The solution is obtained by a reduced rank form of the Procrustes rotation according to Theorem 4 in [8]. We compute the SVD

$$(\mathbf{H}_{(n)}\mathbf{H}_{(n)}^T)\mathbf{B} = \mathbf{UDV}^T$$

and set $\mathbf{A} = \mathbf{UV}^T$.

Given a fixed $\mathbf{A}$, since $\mathbf{A}$ is orthogonal, let $\mathbf{A}_\perp$ be any orthogonal matrix such that $[\mathbf{A}; \mathbf{A}_\perp]$ is $I_n \times I_n$ orthogonal, where $[\mathbf{A}; \mathbf{A}_\perp]$ means to concatenate matrices $\mathbf{A}$ and $\mathbf{A}_\perp$ along row. Then

$$\begin{aligned}
&\sum_{i=1}^{\tilde{I}_n} \| \mathbf{H}_{(n)}(:,i) - \mathbf{AB}^T\mathbf{H}_{(n)}(:,i)\|^2 \\
&= \| \mathbf{H}_{(n)}^T - \mathbf{H}_{(n)}^T\mathbf{BA}^T\|^2 \\
&= \| \mathbf{H}_{(n)}^T\mathbf{A}_\perp\|^2 + \| \mathbf{H}_{(n)}^T\mathbf{A} - \mathbf{H}_{(n)}^T\mathbf{B}\|^2 \\
&= \| \mathbf{H}_{(n)}^T\mathbf{A}_\perp\|^2 + \sum_{p=1}^{P_n} \| \mathbf{H}_{(n)}^T\mathbf{a}_p - \mathbf{H}_{(n)}^T\mathbf{b}_p\|^2
\end{aligned} \tag{6}$$

Because $\mathbf{A}$ is fixed, so the optimal $\mathbf{B}$ minimizing Eq.(5) should minimize:

$$\arg\min_{\mathbf{B}} \sum_{p=1}^{P_n} \| \mathbf{H}_{(n)}^T\mathbf{a}_p - \mathbf{H}_{(n)}^T\mathbf{b}_p\|^2 \tag{7}$$

which is equivalent to P_n independent ridge regression problems. The eigen-decomposition problem is transformed into P_n independent ridge regression problems. However, the ridge regression does not provide a sparse solution. To obtain a sparse solution, Lasso adds an ℓ_1 penalty to the objective function in the regression problem. So Eq.(7) can be transformed to:

$$\mathbf{b}_j = arg\min_{\mathbf{b}_p} \| \mathbf{H}_{(n)}^T\mathbf{a}_p - \mathbf{H}_{(n)}^T\mathbf{b}_p\|^2 + \xi_{1,p}\| \mathbf{b}_p\|_1$$

where $\mathbf{A}^T\mathbf{A} = I$, $p = 1, 2, \dots P_n$. The above equation is the form of elastic net regression problem [8]. So in this paper, Elastic Net is used to obtain the sparse solutions. Due to the nature of the ℓ_1 penalty, some coefficients will be shrunk to zero if $\xi_{1,p}$ is large enough, that is, $\xi_{1,p}$ controls the sparseness.

3. Experiments

We conducted the experiments on two well-known face database Georgia Tech and AR face database. A nearest neighbor classifier based on Manhattan distance is used for recognition.

The sample images of one individual from the Georgia Tech database are shown in Fig. 1.

Figure 1. Sample images on the Georgia Tech.

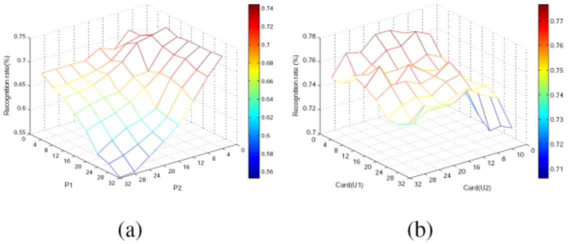

Figure 2. the experiments on Georgia Tech database. (a)the variation of P_1, P_2 and recognition rate of MPCA.(b)the recognition rate of STPCA versus $Card(\mathbf{U}_1)$ and $Card(\mathbf{U}_2)$ when $P_1 = 12$ and $P_2 = 4$.

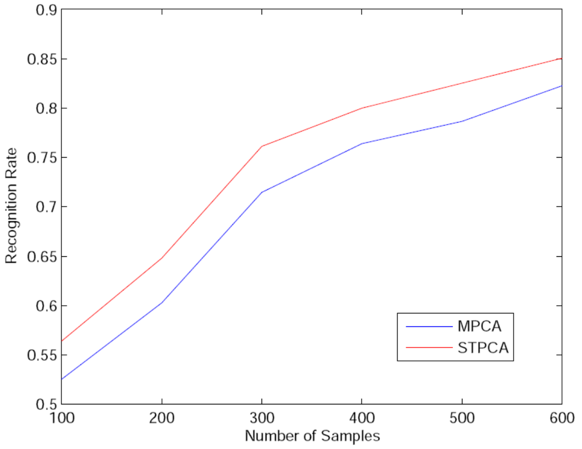

Figure 3. Recognition rate of two methods on Georgia Tech face database.

In this experiment, for both algorithms, the convergence threshold ϵ was set 0.001. It is difficult to determine the optimal dimensionality of the projected subspace. We searched P_1 from 1 to 32 and P_2 from 1 to 32, and selected the projected dimensionality where MPCA had the best performance (see Fig. 2(a)). In order to compare STPCA with MPCA, we set the projected dimensionality of STPCA as the projected dimensionality which MPCA obtained the best performance. Changing the sparseness of projection matrices, different recognition rates were obtained. The recognition rates versus the sparseness are shown in Fig. 2(b). Based on Fig. 2, we set $P_1 = 12$, $P_2 = 4$, $Card(\mathbf{U}_1) = 4$, $Card(\mathbf{U}_2) = 16$ in the following experiment. Each individual's images were divided into 5 bits, and each bit had 3 images. Leave-one-out cross-validation was performed, i.e. for each individual's images, 4 bits were used for training and the remaining bit was used for testing. Fig. 3 is the recognition rates of MPCA and STPCA. From the experiment's result, we can draw a conclusion that STPCA can extract the features of the face images more effectively than MPCA.

We test the robustness of the proposed STPCA. We focus on cases where there are occlusions in the testing set. The experiment was performed in AR face database. All images were cropped into 32×32 pixels, the sample images of one person are shown in Fig. 4. In this experiment, we use the face images without occlusions for training (first row in Fig. 4) and the images with occlusions for testing (second row in Fig. 4).

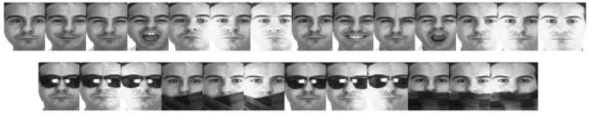

Figure 4. Sample images on the AR database.

The MPCA can achieve its maximal recognition rate 59.17% when the samples are projected into a subspace $\mathbb{R}^{32\times32}$. In order to compare STPCA with MPCA, STPCA also projected the samples into the same dimension. Through changing the sparseness of projection matrices, different recognition rates are obtained. The recognition rates are shown in Table 1 when the samples are projected into a subspace $\mathbb{R}^{32\times32}$. From the table we can see that the algorithm we proposed has a higher recognition rate than MPCA. In MPCA, the whole 2D image is projected to the non-sparse optimal projection $\mathbf{U}_1,\mathbf{U}_2$, i.e $\mathcal{Y}_m = \mathcal{X}_m \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2$, since the elements in $\mathbf{U}_1$ $\mathbf{U}_2$ are non-zero, each pixel of the image matrix is contributive to the feature of $\mathcal{Y}_m$. However, if $\mathbf{U}_1$ $\mathbf{U}_2$ are sparse matrices, only a subset of pixel of the images is contributive to the $\mathcal{Y}_m$. So, STPCA can address the problem of face obscured effectively.

Table 1. the recognition rates on the AR face database

	PCA	MPCA	STPCA
Recognition rate(%)	22.83	59.17	75.33
$\mathbf{U}_1$	1024	32×32	32×32
$\mathbf{U}_2$	-	32×32	32×32
Card($\mathbf{U}_1$), Card($\mathbf{U}_2$)	-	-	16,16

In order to investigate the intuitionistic or semantic interpretation of STPCA, we illustrate the eigentensor representation results in Fig. 5, and eigentensor $\ddot{U}_{p_1p_2} = \mathbf{U}_1(:,p_1)\circ\mathbf{U}_2(:,p_1)$. Based on the projection vectors of each method, facial images can be mapped into each subspace spanned by corresponding a eigentensors. The black points represent the features corresponding to non-zero coefficients of eigentensor. To be more clear, the figure is formed from the non-zero elements in eigentensors, which is then used to construct a mask template, which masks the original face image. From the figure, we can conclude that the areas such as the nose, cheek, and the area around the eyes, mouth and the edges of the facial image, are the main contributors to the new transformed features. For example, the sixth eigentensor in the first row is made up of the black points of the original features, which include the important areas of the cheek, and the area around the nose and mouth. These areas match the conclusion in [3].

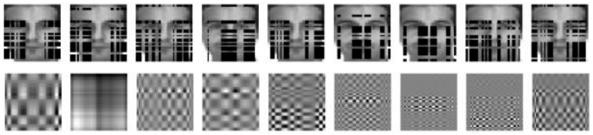

Figure 5. Some eigentensors

References

[1] T. Kolda and B. Bader. Tensor decompositions and applications. SIAM review, 51(3):455–500, 2009.

[2] H. Lu, K. Plataniotis, and A. Venetsanopoulos. MPCA: Multilinear principal component analysis of tensor objects. Neural Networks, IEEE Transactions on, 19(1):18–39, 2008.

[3] O. Ocegueda, S. Shah, and I. Kakadiaris. Which parts of the face give out your identity? In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 641–648. IEEE, 2011.

[4] S.Wang, J. Yang, M. Sun, X. Peng, M. Sun, and C. Zhou. Sparse tensor discriminant color space for face verification. IEEE Transactions on Neural Networks and Learning Systems, 23(6):876 – 888, 2012.

[5] S. Wang, J. Yang, N. Zhang, and C. Zhou. Tensor discriminant color space for face recognition. Image Processing, IEEE Transactions on, 20(9):2490–2501, 2011.

[6] C. Xiao and Z. Wang. Two-dimensional sparse principal component analysis: A new technique for feature extraction. In Natural Computation (ICNC), 2010 Sixth International Conference on, volume 2, pages 976–980. IEEE.

[7] H. Zou and T. Hastie. Regression shrinkage and selection via the elastic net, with applications to microarrays. JR Statist. Soc. B, 2004.

[8] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. Journal of computational and graphical statistics, 15(2):265–286, 2006