Human Action Recognition Using Tensor Principal Component Analysis

Mingfang Sun, Sujing Wang, Xiaohua Liu, Chengcheng Jia, and Chunguang Zhou*
College of Computer Science and Technology
Jilin University
Changchun 130012, China
{sunmf09, wangsj08, jiacc10}@mails.jlu.edu.cn; {xiaohua, cgzhou}@jlu.edu.cn

Abstract—**Human action can be naturally represented as multidimensional arrays known as tensors. In this paper, a simple and efficient algorithm based on tensor subspace learning is proposed for human action recognition. An action is represented as a 3th-order tensor first, then tensor principal component analysis is used to reduce dimensionality and extract the most useful features for action recognition. So the spatial and temporal correlations of the action are preserved. After then, a nearest neighbor classifier based on tensor distance is used to recognize action, in other words, measuring the similarity between actions using tensor distance in tensor subspace. The proposed method is assessed by using a public video database, namely Weizmann human action data sets. Experimental results reveal that the proposed method performs very well on that data sets, and robustness test has been carried out to testify the effectiveness.**

Keywords-Tensor Principal Component; Tensor Distance; Human Action Recognition; Tensor Subspace learning

I. INTRODUCTION

The automatic classification of human action is useful for various applications such as video surveillance, human-computer interfaces, and object-level video summarization and retrieval. As the volume of video data increases, most existing digital video surveillance systems provide the infrastructure only to capture, store, and distribute video while exclusively leaving the task of threat detection to persons. Manual analysis of video is labor intensive, fatiguing, and prone to errors. Software-aided real-time video analytics or forensics would considerably alleviate the people constraints, which currently are the main handicap for analyzing continuous surveillance data, with focusing on surveillance of human action, human behavior recognition has become one of the most active research topics in image processing and pattern recognition.

With gradually deepening the study on action recognition, a variety of methods have been proposed to construct action classifier [1-3], these human action recognition methods can basically be divided into two classes. One is based on motion features, and the other is based on shapes features. The recognition method, based on motion features is to represent and identify human behavior by extracting the information of their movement characteristics, such as movement direction, speed, optical flow information. Zhu *et al.* [4] represent the basic player action in broadcast tennis video by a group of histograms based on optical flow, and employ Support Vector Machine to train the classifier for recognition the action. Even in long distances or poor visibility, the method is still be able to identify the type of target motion according to human action mode easily, but it is difficult to extract the motion feature of target from video accurately with current computer vision technology. The recognition method, based on shape features, is to represent and identify human behavior by recovering geometric information of human posture, such as silhouettes, profile and so on. Li *et al.* [5] exploit the changes in contours along the spatio-temporal direction, each contour is first parameterized as a 2D function of radius and a 3D surface is composed from a sequence of such functions, and they employ Dynamic Time Warping and Mutual Information to recognize human action. The comparison of two video volumes has been achieved by matching templates called motion history images [6]. The shape feature is often obscure in long distances or poor visibility, but they are easy for available and not sensitive to the texture. Therefore, shape feature has been widely applied in human action recognition.

The tensor has been used in recognition area such as face recognition and human action recognition more and more frequently. Wang *et al.* [7] propose Discriminant Tensor Subspace Analysis (DTSA) algorithm to extract discriminant features from the intrinsic manifold structure of the 2^{nd}-order tensor. They also treat a color facial image as a 3^{rd}-order tensor and propose tensor discriminant color space (TDCS) mode [8] to seek an optimal color space for face recognition. Zhou *et al.* [9] transform space-time human silhouettes extracted from action sequences to a low dimensional multivariate time series using tensor subspace analysis, then employ Gaussian Processes classification to learn and predict action categories. Feature vectors are extracted from silhouette images of action sequences and the Euclidian distance of the feature vectors is served as similarity of action sequences [10]. They extract features using frame-by-frame processing and did not fully utilize the nature spatial and temporal correlations of the input data. Wu *et al.* [11] converted silhouettes into vectors and described an action sequence with a matrix, the methods destroy the nature spatial and temporal correlations of the original action data. Lu *et al.* [12] propose Multilinear Principal Component Analysis (MPCA) to capture most

* Corresponding author.

variance for tensor dimensionality reduction, and use it for gait recognition [13]. Chen *et al.* [14] use MPCA for face recognition.

In the recognition method based on shape features, the spatial and temporal variability of the silhouettes reflects the human action information, that is to say, the spatial and temporal relationship between pixels of the result after a series' transformations on silhouette sequence, will affect the recognition result. So the correct recognition rate will be improved if spatial and temporal information is effectively maintained while reducing the dimension. Inspired by prior work, an action sequence is represented as a *3*th order tensor in this paper, and be projected into low-dimension subspace by tensor subspace learning. In the process of reducing the dimension, MPCA is used to extract the features. After dimension reduced, the tensor distance [15] is used to the measurement to construct a nearest neighbor classifier for action classification.

The rest of this paper is organized as follows: Tensor algebra is reviewed in Section 2. In section 3, tensor subspace learning and action classification are presented. Experimental results are reported in Section 4. Finally, conclusions are given in Section 5.

II. MATHEMATICAL BACKGROUND

A tensor is a multidimensional array. More formally, an *N*th-order tensor is an element of the tensor product of *N* vector spaces. The order of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is *N*. An element of $\mathcal{A}$ denoted by $\mathcal{A}_{i_1 i_2 \ldots i_N}$, where $1 \le i_n \le I_n, n = 1,2,\ldots N$. The n-mode unfolding matrix of $\mathcal{A}$, denoted by $\mathbf{A}_{(n)} \in \mathbb{R}^{I_n \times (I_1 \times \ldots \times I_{n-1} \times I_{n+1} \times \ldots \times I_N)}$. The pictorial description is given in Fig. 1 for a 3th-order tensor.

The n-mode product of a tensor of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{U}$, is an $(I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N)$-tensor of which the entries are given by:

$$(\mathcal{A} \times_n \mathbf{U})_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} \stackrel{\text{def}}{=} \sum_{i_n} a_{i_1 i_2 \ldots i_{n-1} i_n i_{n+1} \ldots i_N} u_{j_n i_n} \quad (1)$$

By using tensor decomposition, any tensor $\mathcal{A}$ can be expressed as the product:

$$\mathcal{A} = \mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \ldots \times_N \mathbf{U}_N \quad (2)$$

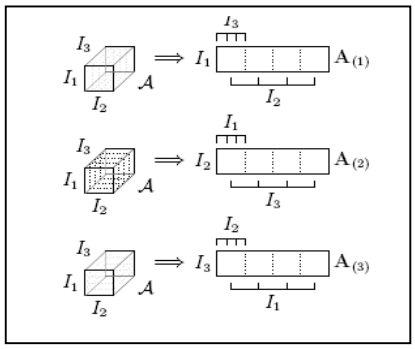

Figure 1. An example of matrix unfolding of a 3th-order tensor

Where $\mathcal{S} = \mathcal{A} \times_1 \mathbf{U}^{(1)T} \times_2 \mathbf{U}^{(2)T} \times \ldots \times_N \mathbf{U}^{(N)T}$ is called the core tensor, $\mathbf{U}_n, n = 1,2,\ldots,N$, is an orthogonal matrix and contains the ordered principal components for the n-th mode.

III. TENSOR SUBSPACE LEARNING AND CLASSIFICATION

Human action recognition using tensor principal component analysis, is to project action sequences into low-dimensional tensor subspace for recognizing action. As shown in Fig. 2.

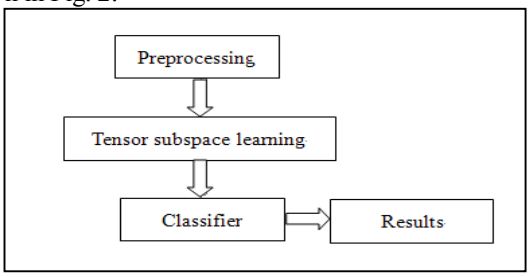

Figure 2. Principal diagram of human action recognition

Based on the statistical theory, tensor subspace learning is to solve the most contributing tensors set for optimization goal, according to some definition of the optimization criterion. And constitute a tensor subspace by this tensors set. When classification, firstly, get the corresponding projection coefficient by mapping the training samples and test samples to the subspace, Secondly, calculating the similarity between the test sample and pattern class based on the specific metrics. Finally, determine the type of the test samples. Research shows that action space is a low dimensional subspace embedded in a high one, the dimension and the corresponding tensors set of the subspace reflect the difference between human action classes. Feature extraction is to find the essence dimension and the corresponding tensors set of the action space.

A. Tensor subspace learning

In the process of constructing a tensor subspace, MPCA algorithm is used to generate projection matrices and determine the dimension of the subspace, and hope it can achieve a high recognition rate.

MPCA(Multilinear Principal Component Analysis) can use matrices directly to capture most variance for dimensionality reduction. An action sequence is naturally represented by a 3th-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the 1-mode of tensor is the height of silhouette image in the sequence, the 2-mode of tensor is the width of silhouette images and the 3-mode of tensor is the number of frames of the action sequence. A set of M 3th-order tensor action sequences $\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_M\}$ is available for training. MPCA algorithm is to seek 3 projection matrices $\mathbf{U}_1 \in \mathbb{R}^{P_1 \times I_1}$, $\mathbf{U}_2 \in \mathbb{R}^{P_2 \times I_2}$, $\mathbf{U}_3 \in \mathbb{R}^{P_3 \times I_3}$, (usually $P_1 < I_1$, $P_2 < I_2$, $P_3 < I_3$) to transform:

$$\mathcal{Y}_i = \mathcal{X}_i \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \tag{3}$$

And meeting that they can maximize the total tensor scatter Ψ:

$$\{\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3\} = arg \max_{\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3} \Psi \tag{4}$$

Where $\Psi = \left\| \overline{\mathcal{Y}_i} \right\|_F^2, 1 \le i \le N$.

For computing $\{\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3\}$, firstly, we set $\mathbf{U}_i$ to consist of the eigenvectors corresponding to the most P_n eigenvalues of $\sum_{m=1}^{M} \mathbf{X}_{m(n)} \cdot \mathbf{X}_{m(n)}^{T}$. Then update the matrix $\mathbf{U}_1$ with the P_1 eigenvector corresponding to the largest P_1 eigenvalues of the matrix:

$$\Phi^{(1)} = \sum_{m=1}^{M} (\mathcal{X}_{m(1)} - \overline{\mathcal{X}_{(1)}})(\mathbf{U}_2 \otimes \mathbf{U}_3)(\mathbf{U}_2 \otimes \mathbf{U}_3)^T (\mathcal{X}_{m(1)} - \overline{\mathcal{X}_{(1)}})^T \tag{5}$$

where $\otimes$ is the Kronecker product of the matrix. $\mathbf{U}_2$ and $\mathbf{U}_3$ are updated by using the same principle. The process is repeated until the convergence.

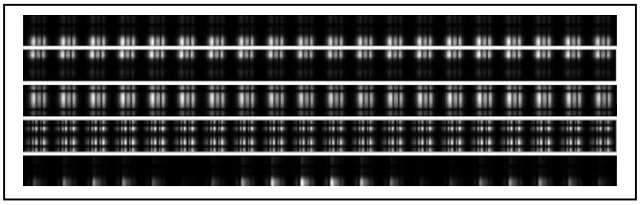

Figure 3. The 1-mode unfolding of the first, second, third, 40th, and 100th EigenActions

Each tensor $\mathcal{X}_i$ can be viewed as a linear combination of $P_1 \times P_2 \times P_3$ EigenActions, with each entry of projected tensor $\mathcal{Y}_i$ corresponding to one EigenAction. A part of the 1-mode unfolding of EigenActions is shown in Fig. 3.

B. *Nearest Neighbor Classifier based on Tensor Distance*

Each training action sequence $\mathcal{X}_i$ should be projected onto tensor subspace $\mathbb{R}^{P_1 \times P_2 \times P_3}$ using (3) to yield a tensor $\mathcal{Y}_i$. Given a test action sequence $\mathcal{X}_{test}$, the corresponding $\mathcal{Y}_{test}$ is:

$$\mathcal{Y}_{test} = \mathcal{X}_{test} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \tag{6}$$

For data after dimensionality reduction, a nearest neighbor classifier based on tensor distance is constructed.

$$i^* = arg \min_{i \in M} d_{TD}(\mathcal{Y}_i, \mathcal{Y}_{test}) \tag{7}$$

Given a 3th-order $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, $\mathbf{x}$ is used to denote the vector form representation of $\mathcal{X}$, the element $\mathcal{X}_{i_1 i_2 i_3}$ in $\mathcal{X}$ is corresponding to $\mathbf{x}_l$, the lth element in $\mathbf{x}$, where $l = i_1 + (i_2 - 1) \times I_1 + (i_3 - 1) \times I_1 \times I_2$, the element $\mathcal{X}_{i_{1'} i_{2'} i_{3'}}$ is corresponding to $\mathbf{x}_m$, $m = i_{1'} + (i_{2'} - 1) \times I_{1'} + (i_{3'} - 1) \times I_1 \times I_2$ then

$$d_{TD}(\mathcal{X}, \mathcal{Y}) = \sum_{l,m=1}^{P_1 \times P_2 \times P_3} g_{lm} (\mathbf{x}_l - \mathbf{y}_m)(\mathbf{x}_l - \mathbf{y}_m) \tag{8}$$

where g_{lm} is metric coefficient, and

$$g_{lm} = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{(i_1 - i_{1'})^2 + (i_2 - i_{2'})^2 + (i_3 - i_{3'})^2}{2\sigma^2} \right\} \tag{9}$$

IV. EXPERIMENTAL RESULTS

The experiment are conducted to evaluate the proposed method on Weizmann human action dataset [11, 16-17]. In the dataset, there are about 90 low-resolution video sequences showing nine different people performing 10 natural actions such as bending(bend), jumping jack(jack), jumping-forward-on-two-legs(jump), jumping-in-place-on-two-legs (pjump), running(run), skipping(skip), galloping-sideways (side), walking (walk), waving-one-hand(wave1) and waving -two-hand(wave2). Examples of these actions are given in Fig. 4.

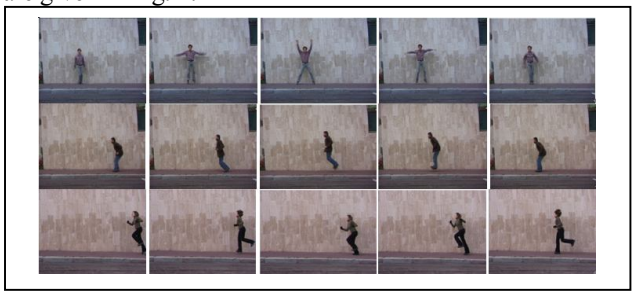

Figure 4. Rows from top to bottom are examples of jacking, jumping and skipping

TABLE I. CONFUSION MATRICE FOR HUMAN ACTION CLASSIFICATION

	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	1.00	0	0	0	0	0	0	0	0	0
Jack	0	1.00	0	0	0	0	0	0	0	0
Jump	0	0	1.00	0	0	0	0	0	0	0
Pjump	0	0	0	1.00	0	0	0	0	0	0
Run	0	0	0	0	0.98	0	0	0.02	0	0
Side	0	0	0	0	0	0.98	0	0.02	0	0
Skip	0	0	0	0	0	0	1.00	0	0	0
Walk	0	0	0	0	0	0	0	1.00	0	0
Wave1	0	0	0	0.03	0	0	0	0	0.97	0
Wave2	0	0	0	0	0	0	0	0	0	1.00

In the process of extract tensor subspace features using MPCA alagorithm, the targeted dimension $\{P_n, n=1,2,3\}$ is not specified,but decided by the percentage of variation(testQ) kept in each mode, where testQ= $\left\{\sum_{i_n=1}^{P_n} \lambda_{i_n}^{(n)*}\right\} / \left\{\sum_{i_n=1}^{I_n} \lambda_{i_n}^{(n)*}\right\}$. With the testQ,the first P_n eigenvalue of the total scatter are kept in the n-mode. With the different of testQ, the recognition rates are different. The recognition rates with different testQ are showed in Fig. 5. From the figure we can learn that reserving too much or too little eigenvalues is not good for recognition rate. In the remaining experiment, we set testQ=80.

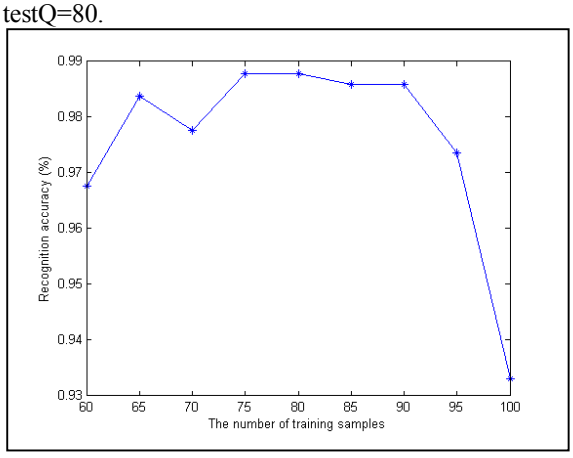

Figure 5. The recognition rates with different testQ.

The classification accuracy is evaluated under nine-fold cross validation. In each fold, take the actions of eight people for training and use those of the remaining one for testing. For each action sequence of a person in the database, it's centered silhouettes extracted in [16]. 20 frames are taken interval five frames as a small action. In this case, an action sequence in the database can be divided into lots of small actions, which size is 64×48×20. Training set or testing set is composed by those small actions. Furthermore, no space-time alignment is performed on this data set. The results are illustrated in the Table 1. From the table, we can see that: a higher recognition rate can be achieved with our method.

To evaluate the adaptability and robustness of our method for action recognition to the irregular actions in changing scenarios, ten test video sequences of people walking in various difficult scenarios in front of different non-uniform backgrounds are collected, some example frames are illustrated in Fig 6.

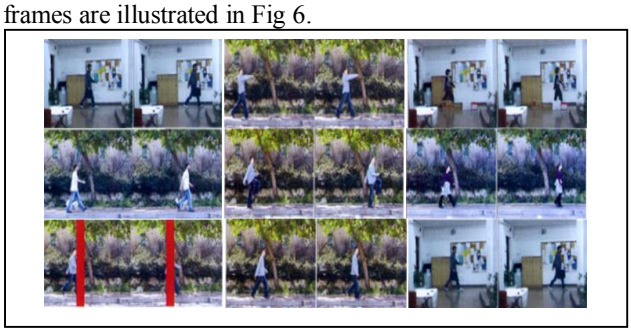

Figure 6. Test walking sequences under different conditions. Rows from top to bottom, lines from left to right are examples of walking with swimming a bag, moonwalk, walking with occluded feet, walking with a dog, walking with knees up, walking in a skirt, walking occluded by a pole, limp walking and walking with a briefcase.

TABLE II. ROBUSTNESS TEST RESULTS.

	Jack	Side	Skip	Walk
Swinging a bag<8>	0	2	0	6
Carry a briefcase<5>	0	0	0	5
Limp walk<18>	2	5	1	10
Moonwalk<10>	0	0	0	10
Walk in a skirt<7>	0	0	1	6
Walk with a dog<7>	0	0	0	7
Occluded feet<8>	0	0	0	8

In this experiment, 10 actions (bend, jack, jump, pjump, run, skip, side, walk, wave1, wave2) of nine people are used as training set. An irregular action is divided into actions as the above described. The result of recognition is shown in Table 2. The number in parentheses, for example <18>, indicates that the behavior of limp walking is divided into 18 small actions for recognition. From the table, we can know that irregular actions can be recognized accurately except walking with knees up.

An action sequence is represented as a 3th-order tensor. In this way, the spatial and temporal structure information of action sequences can be maintained in the process of tensor

subspace learning. MPCA performs feature extraction by determining a multilinear projection that captures most of the original tensorial input variation, From the (4), the variance (Ψ)projected by MPCA is maximized, in this way the subspace generated by MPCA is the smallest reconstruction subspace, and the Ψ greater, the more information provided, whereas the less information provided. The noise in the action sequences, treated as corresponding to the small eigenvalues, are removed, in other words, data after dimensionality reduction retain most useful information of resource data for recognition. Tensor distance does not ignore the relationships among different coordinates for high-order data, so it can reflect the real distance between two data points.

V. CONCLUSIONS

In this paper, unlike transforming an action sequence into a matrix, by which spatial and temporal correlations of the action are broken in tensor subspace learning, an action sequence is represented as a 3th-order tensor. We reduce dimensionality effective through multilinear principal component analysis, at the same time, spatial and temporal correlations of action sequence are preserved. Experiments on both regular and irregular actions shown that the proposed method can reach high accuracy for recognition with small number of training samples. Unlike the method from [11] which uses thousands training samples, only 600-800 training samples are used in our experiments. We will extend our research on how to recognize a new action that does not exist in the training set as a new action class in our future work.

ACKNOWLEDGMENT

This paper is supported by (1) the National Natural Science Foundation of China under Grant No.60873146, 60973092,60903097, (2) National High Technology Research and Development Program of China under Grant No. 2007AAO4Z114, (3) project of science and technology innovation platform of computing and software science (985 engineering).

REFERENCES

[1] G. Xu and Y. Cao, "Action recognition and activity understanding: A review," Journal of Image and Graphics, vol. 14, pp. 189-195, 2009.

[2] J. Gu, X. Ding, and S. Wang, "A survey of activity analysis algorithms," Journal of Image and Graphics, vol. 14, pp. 377-387, 2009.

[3] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi, "Understanding transit scenes: a survey on human behavior-recognition algorithms," Intelligent Transportation Systems, IEEE Transactions on, vol. 11, no. 1, pp. 206–224, 2010.

[4] G. Zhu, C. Xu, Q. Huang, and W. Gao, "Action recognition in broadcast tennis video," in Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 1. IEEE, 2006, pp. 251–254.

[5] H. Li and M. Greenspan, "Multi-scale gesture recognition from timevarying contours," in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 1. IEEE, 2005, pp. 236–243.

[6] M. Yang, F. Lv, W. Xu, K. Yu, and Y. Gong, "Human action detection by boosting efficient motion features," in Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. IEEE, 2010, pp. 522–529.

[7] S. Wang, C. Zhou, N. Zhang, X. Peng, Y. Chen, X. Liu, "Face recognition using second order discriminant tensor subspace," Neurocomputing, in press.

[8] S. Wang, J. Yang, N. Zhang, and C. Zhou, "Tensor Discriminant Color Space for Face Recognition," IEEE transaction on image processing: a publication of the IEEE Signal Processing Society, doi: 10.1109/TP.2011.2121084, in press.

[9] H. Zhou, L. Wang, and D. Suter, "Human action recognition by feature-reduced Gaussian process classification," Pattern Recognition Letters, vol. 30, no. 12, pp. 1059–1066, 2009.

[10] Z. Ling, Y. Liang, Q. Pan, Y. Cheng, and Z. C., "Human action recognition based on tensor subspace learning." Journal of Image and Graphics, vol. 14, pp. 394–400, 2009.

[11] X. Wu, W. Liang, and Y. Jia, "Incremental discriminative-analysis of canonical correlations for action recognition," in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2010, pp. 2035–2041.

[12] H. Lu, K. Plataniotis, and A. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," Neural Networks, IEEE Transactions on, vol. 19, no. 1, pp. 18–39, 2008.

[13] H. Lu, K. Plataniotis, and A. Venetsanopoulos, "Gait recognition through MPCA plus LDA," in Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research at the. IEEE, 2007, pp. 1–6.

[14] C. Chen, Z. Shi-qing, and C. Yue-fen, "Face recognition based on MPCA," in Industrial Mechatronics and Automation (ICIMA), 2010 2nd International Conference on, vol. 1. IEEE, 2010, pp. 322–325.

[15] Y. Liu and K. Chan, "Tensor Distance Based Multilinear Locality-Preserved Maximum Information Embedding." IEEE transactions on neural networks/a publication of the IEEE Neural Networks Council, 2010.

[16] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," 2005.

[17] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html#Classification%20Database