PAPER Special Section on Machine Vision and its Applications

View Invariant Human Action Recognition Based on Factorization and HMMs

Xi LI[†a)], *Nonmember and* Kazuhiro FUKUI[†], *Member*

SUMMARY This paper addresses the problem of view invariant action recognition using 2D trajectories of landmark points on human body. It is a challenging task since for a specific action category, the 2D observations of different instances might be extremely different due to varying viewpoint and changes in speed. By assuming that the execution of an action can be approximated by dynamic linear combination of a set of basis shapes, a novel view invariant human action recognition method is proposed based on non-rigid matrix factorization and Hidden Markov Models (HMMs). We show that the low dimensional weight coefficients of basis shapes by measurement matrix non-rigid factorization contain the key information for action recognition regardless of the viewpoint changing. Based on the extracted discriminative features, the HMMs is used for temporal dynamic modeling and robust action classification. The proposed method is tested using real life sequences and promising performance is achieved.

key words: *action recognition, view invariant, matrix factorization, Hidden Markov Models*

1. Introduction

Recently, human action recognition has become an active research area due to its many potential applications such as video surveillance, human-computer interface, robot maneuvering, content based video retrieval and sports video analysis.

Many approaches for human action recognition have been proposed previously. The most common one taken by the researchers is to perform action recognition using 2D observation, such as silhouettes of the target subject or landmark points trajectories. For example, temporal template was used for human movement representation and recognition [1], where the temporal template was a static vector-image and the vector value at each point was a function of the motion properties at the corresponding spatial location in an image sequence. This method can perform temporal segmentation automatically and is invariant to linear changes in speed. The motion descriptor based on optical flow measurements in a spatial-temporal volume for each stabilized human figure was later introduced and an associated similarity measurement was used in a nearest-neighbor framework [2]. Cubic higher order local auto-correlation (CHLAC), the extension of the traditional higher order local auto-correlation (HLAC) to the three-way data analysis case, was exploited for action and simultaneous multiple-person identification [3]. A sparse representation of image sequences as a collection of spatiotemporal events that are localized at detected spatiotemporal salient points was proposed later in [4], [5]. Based on novelly defined distance metric and relevance vector machines, promising classification result was achieved on a 19 aerobic exercises database. A real time system for recognition of 15 different continuous human activities was presented in [6], where the actions were represented as a continuous sequence of discrete postures which were derived from affine invariant descriptor. Both of the above methods are viewpoint dependent. That is to say, the training sequences and testing sequences are captured under the same viewing direction by stationary cameras. But in real life applications, the training sequences and testing sequences are not necessarily captured from the same viewpoint. The 2D observations of different action instances might be extremely different even if correspond to the same action category. The situation is even worse if the sequences are captured by a moving camera and the viewpoints are varying on-the-fly. Furthermore, usually the actions are executed at different rates, which renders the problem much harder.

Several view invariant human action recognition methods have been proposed. For example, a computational representation of human action using spatial-temporal curvature of 2D trajectory was presented in [7]. [8] proposed a 3D model based view invariant human action recognition method. The epipolar geometric constraints computed from the correspondences of human body landmarks were used to match actions performed from different viewpoints and in different environments [9]. In [10], the human action was represented by a set of descriptor computed from a spatial-temporal action volume created from a set of object silhouette. Again, the epipolar geometry between the views of two stationary cameras was exploited to achieve view invariant recognition. The above view invariant action recognition methods have the limitation that action sequences are captured using stationary cameras. The traditional epipolar geometry was further extended to the geometry of dynamic scenes where the projection camera was no longer stationary [11].

In this paper, using the trajectories of landmark points on the human body as inputs, a simple yet effective view invariant human action recognition method is proposed based on non-rigid factorization and Hidden Markov Models (HMMs). By assuming that the execution of a specific action can be approximated by dynamic linear combination of a set of basis shapes, we show that the low dimen-

Manuscript received November 5, 2007.
Manuscript revised January 25, 2008.

[†]The authors are with Computer Vision Laboratory, Tsukuba University, Tsukuba-shi, 305–8573 Japan.
a) E-mail: lixitsukuba@gmail.com

DOI: 10.1093/ietisy/e91–d.7.1848

sional weight coefficients of basis shapes by measurement matrix non-rigid factorization contain the key information for action recognition regardless of the viewpoint changing. Based on the extracted discriminative features, the HMMs, which allows for the inclusion of temporal dynamics, is used for action modeling and classification. The proposed method is tested using real life sequences and promising performance is achieved.

The rest of this paper is organized as follows: Section 2 describes the discriminative feature extraction based on non-rigid factorization. Section 3 presents the method of applying HMMs to human action modeling and recognition after a brief review of HMMs. Experimental results using real life database are presented in Sect. 4, followed by conclusions in Sect. 5.

2. Feature Extraction Based on Non-rigid Factorization

As in [8]–[11], this work does not address the lower-level processing tasks such as body-joint detection and tracking. We concentrate on how to construct discriminative features for action recognition under varying viewpoint directions and different execution speed, given the 2D trajectories of anatomical landmarks on human body. There are many possible sets of features that could be used for action recognition, but the optimal choice for view invariant recognition is not obvious. It is difficult to recognize actions, either captured by stationary cameras with different viewpoint or by moving on-the-fly cameras, because the 2D observations might look quite different even the same person performing action of the same category. This is true both for contour based representations and landmark trajectories based representations. It is demonstrated in Fig. 1 using sample sequences for walking, running and jumping. Taking the walking sequences for example, Figs. 1 (a) and (d) are two walking sequences performed by different persons. Figures 1 (b) and (e) are the 2D trajectories observations for the two walking sequences under same viewing directions by stationary camera, respectively. Even the two sequences are performed by different persons, the 2D observations still look similar since they belong to the same action category and the body joints move in a similar way. Figures 1 (c) and (f) are the 2D trajectories observations for the two walking sequences projected using moving cameras, with the trajectories superimposed. It can be clearly seen that, due to the motion of the camera, not only the trajectories in Figs. 1 (c) and (f) do not appear similar, but also the trajectories pairs in Figs. 1 (b) (c) and (e) (f) look quite different even these sequences, which belong to the same action category, are performed by the same person. Figures 1 (g)–(l) and (m)–(r) show the examples for running and jumping case respectively and the same conclusion can be drawn.

Our view invariant approach for human action recognition in videos acquired by non-stationary cameras is based on the observation that a deformable shape, e.g. human body, can be approximately represented by a linear combination of basis shapes, where the weight coefficients assigned to each basis shape change with time. We show that the low dimensional deformation coefficients of basis shapes contain the key information for action recognition regardless of the viewpoint changing. It is well known that both shape and motion can be factorized directly from the measurement matrix constructed from feature point trajectories under orthographic camera model and rigidity assumption [12]. The problem in the action recognition scenario is more complex because the freedom of moving human body is extremely high due to the non-rigidity. The traditional rigid factorization algorithm was further extended to the non-rigid case in [13], [14]. Suppose that P feature landmark points are tracked across F frames, the deforming shape can be described as a key frame basis set $S_1, S_2, \ldots, S_K$, where each key frame S_i is a $3 \times P$ matrix. The shape of a specific configuration is a linear combination of the basis set as follows:

$$S = \sum_{l=1}^{K} l_i S_i, \qquad S, S_i \in R^{3\times P}, l_i \in R \tag{1}$$

When the size of the target subject is relatively small enough compared with the distance between target subject and viewing camera, the projection procedure can be approximated using orthographic model:

$$\begin{bmatrix} u_1 & u_2 & \ldots & u_P \\ v_1 & v_2 & \ldots & v_P \end{bmatrix} = R\left(\sum_{i=1}^{K} l_i S_i\right) + T \tag{2}$$

(u_i, v_i) represents the 2D projection observations of the feature point i. R contains the first two rows of the full 3D camera rotation matrix and T contains the first two components of the camera translation vector. Equation (2) can be rewritten as follows after eliminating T by subtracting the mean of feature points as in [12]:

$$\begin{bmatrix} u_1 & u_2 & \ldots & u_P \\ v_1 & v_2 & \ldots & v_P \end{bmatrix} = \begin{bmatrix} l_1 R & l_2 R & \ldots & l_K R \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_K \end{bmatrix} \tag{3}$$

If we write all the feature points along the temporal axis into a $2F \times P$ measurement matrix W as follows:

$$W = \begin{bmatrix} u_1^1 & u_2^1 & \ldots & u_P^1 \\ v_1^1 & v_2^1 & \ldots & v_P^1 \\ u_1^2 & u_2^2 & \ldots & u_P^2 \\ v_1^2 & v_2^2 & \ldots & v_P^2 \\ \vdots & \vdots & \vdots & \vdots \\ u_1^F & u_2^F & \ldots & u_P^F \\ v_1^F & v_2^F & \ldots & v_P^F \end{bmatrix} \tag{4}$$

then W can be further decomposed into the following form:

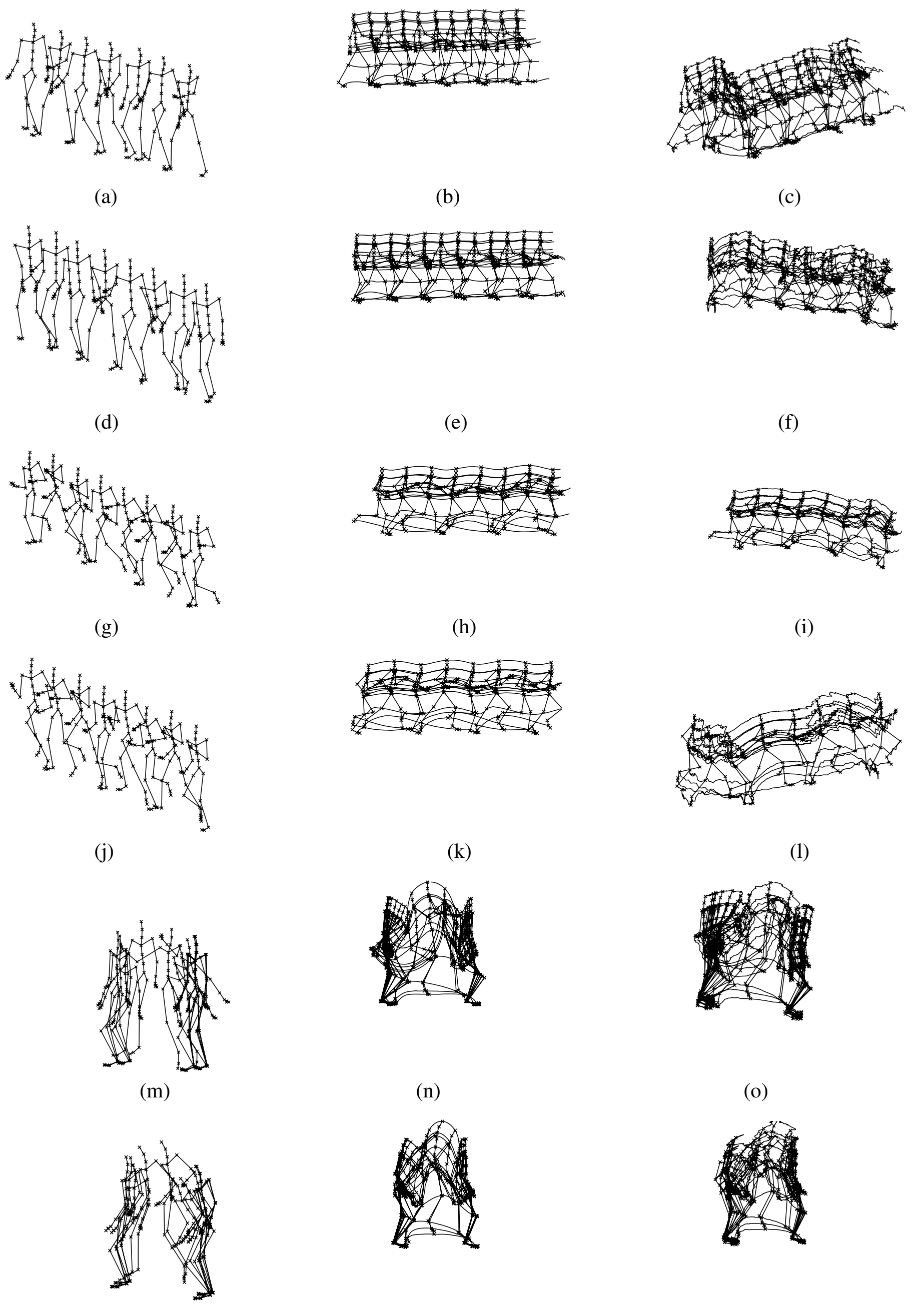


Fig. 1 Example sequences. (a) (d): 3D walking sequences; (b) (e): 2D landmark points trajectory superimposition for walking sequences (a) (d) projected using stationary cameras; (c) (f): 2D landmark points trajectory superimposition for walking sequences (a) (d) projected using non-stationary cameras; for two different performers respectively. (g)–(l) are example sequences for running action and (m)–(r) are example sequences for jumping action. See the main text for detail.

$$W = \begin{bmatrix} l_1^1 R^1 & l_2^1 R^1 & \dots & l_K^1 R^1 \\ l_1^2 R^2 & l_2^2 R^2 & \dots & l_K^2 R^2 \\ \vdots & \vdots & \vdots & \vdots \\ l_1^F R^F & l_2^F R^F & \dots & l_K^F R^F \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_K \end{bmatrix} \tag{5}$$

An effective way for factorization of the measurement matrix W as Eq. (5) was proposed in [13], [14]. Firstly, the weighting coefficients l_k^t, $k = 1, \dots, K$, $t = 1, \dots, F$ are randomly initialized, and then the shape bases S_i, $i = 1, \dots, K$ are computed in the least-square-fit sense. Given an initial guess of the rotation matrix R and the shape basis, the coefficients l can also be solved using linear least squares. Next, given the shape basis and the weight coefficients, the rotation matrix R can be recovered by parameterized with exponential coordinates. The above procedures are iterated until convergence. More details can be found in literature [13],

[14].

Denote the weight coefficient vector corresponding to frame i as $L^{(i)} = (l_1^i, l_2^i, \ldots, l_K^i)$, then the vector sequence $\phi = (L^{(1)}, L^{(2)}, \ldots, L^{(F)})$ describes the changing mode for the body-parts and contains the necessary information for action recognition regardless of viewpoint changing. The reason is that each instance of a specific action class is composed of a sequence of deforming body. The body shape deforms along the time axis while each shape at a specific time stamp is a combination of different basis shapes with different weights on the basis of average shape. The 2D observations of those body shape are view variant. They are quite dependent on the viewpoint and the chosen coordinate system. But the weight coefficient vector is invariant to these conditions. The inherent weight coefficient vector describes the style of the whole body deforming by depicting the changing mode of each relative displacement of the body, which are characterized by the basis shapes. The ϕs for different action categories should exhibit different patterns while the ϕs for same action should have similar patterns, regardless of different subject performers, changing viewpoints or moving on-the-fly capturing cameras. It should be noted that the vector sequence ϕ can not be used directly for action recognition, because in the iteration procedure of the non-rigid factorization, no constraints has been imposed on the shape basis. For action sequences of different instances, the shape basis yield by non-rigid factorization of the measurement matrix might also be different. In order to make the comparison reasonable, we should put the weight coefficients sequences under the same conditions, i.e., they should correspond to the same shape basis set.

Suppose there are C action classes to be recognized. The number of training sequences for the i-th action class is N_i. Denote the measurement matrix for the j-th sequence of the i-th action class as W_i^j, we stack all training sequences vertically as follows:

$$w = \left[W_1^{1^T}, \ldots, W_1^{N_1^T}, \ldots, W_C^{1^T}, \ldots, W_C^{N_C^T} \right]^T \tag{6}$$

Here, we make use of the fact that all human figures share the same skeleton structure. The procedure of stacking measurement matrix can be imagined that the subject undergoes a virtual movement from the position in the last frame of i-th sequences to the position of the first frame in the $(i+1)$-th sequence. After non-rigid factorization, we can get the weight coefficient vector sequences along the temporal axis as ϕ_i^j, $i = 1, \ldots, C$, $j = 1, \ldots, N_i$. If the length of the j-th sequence of the i-th action class is F_i^j, ϕ_i^j can be written in the following form,

$$\phi_i^j = \begin{bmatrix} l_{i(1)}^{j(1)} & l_{i(2)}^{j(1)} & \cdots & l_{i(K)}^{j(1)} \\ l_{i(1)}^{j(2)} & l_{i(2)}^{j(2)} & \cdots & l_{i(K)}^{j(2)} \\ \vdots & \vdots & \vdots & \vdots \\ l_{i(1)}^{j(F_i^j)} & l_{i(2)}^{j(F_i^j)} & \cdots & l_{i(K)}^{j(F_i^j)} \end{bmatrix} \tag{7}$$

Since the different actions share the same shape basis, the discriminative information for action recognition are encoded in the ϕ_i^js. Figures 2 (a)–(c), (d)–(f) and (g)–(i) show the examples of the recovered weight coefficients for action categories of walking, running and jumping. For each action category, the left sub-figure and middle sub-figure are recovered weight coefficients for sequences of two different performers with stationary cameras under the same viewpoint while the right sub-figure is the recovered weight coefficients for sequences from a moving on-the-fly camera. It can be seen that the varying patterns of the weight coefficients look similar for the same action classes, even with different performers or captured with a moving camera. On the other hand, the patterns look quite different for different action classes. Thus the weight coefficients are appropriate for view invariant action recognition of human body under the condition of variability such as captured by moving projection cameras.

3. HMMs Based Action Modeling and Recognition

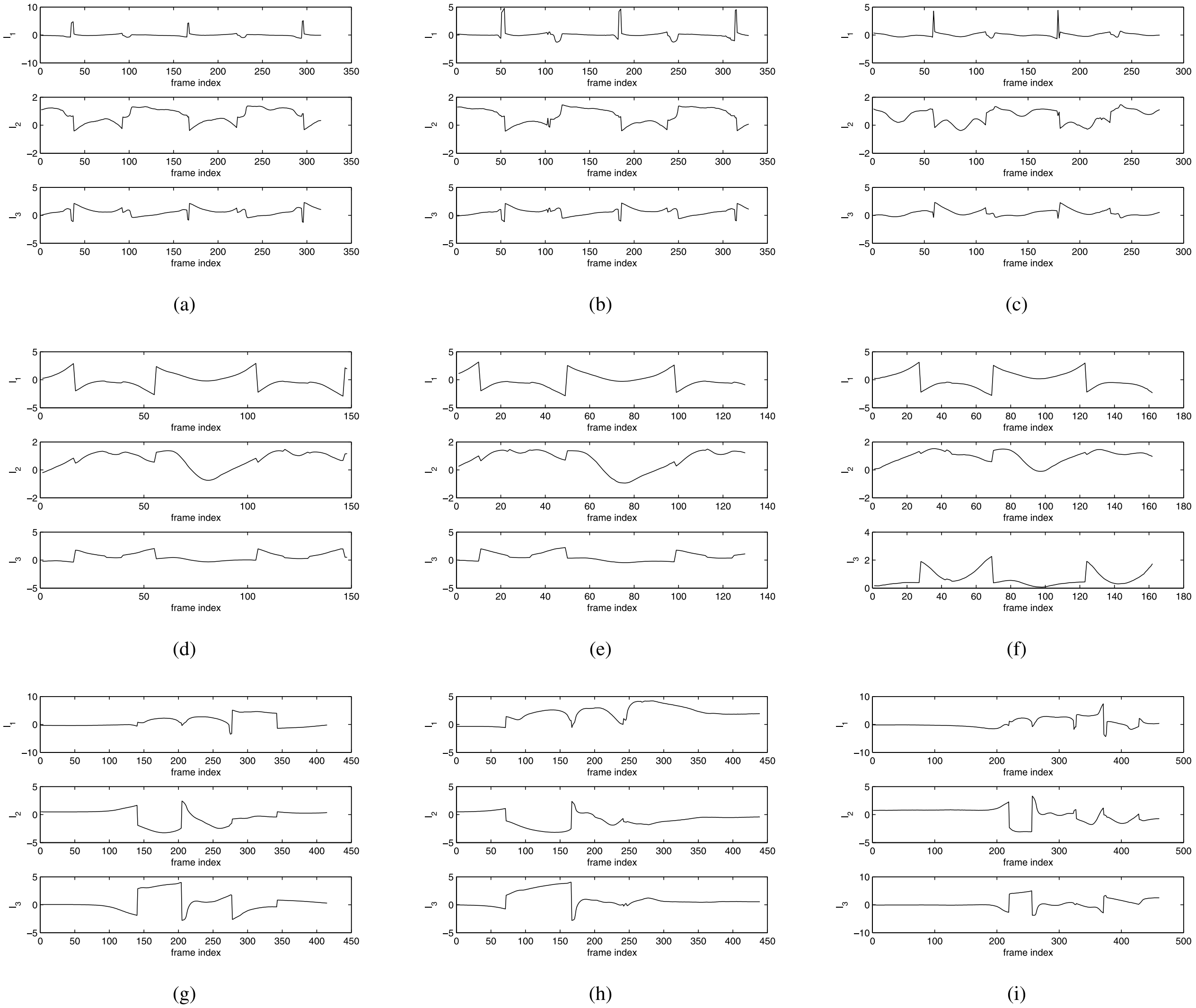
Hidden Markov Models (HMMs) [15] have been successfully used for speech recognition and computer vision. We employ the HMMs for action modeling and recognition since it can be applied to model the time series data well, such as the weight coefficients with temporal variations. It allows for the inclusion of temporal dynamics to model the action sequences. The HMM model for the c-th action class is given by $\lambda_c = (A_c, B_c, \pi_c)$ with N number of states. Here A_c is the transition matrix and π_c is the initial distribution. The B_c parameter represents the probability distributions for the observed feature vector conditional on the hidden states. In this work the HMMs with mixture of Gaussians is used for action modeling. Suppose each state has a bank of M Gaussian components, then the parameter B_c consists of the following items: the mean vector μ_{im}, the mixture coefficient c_{im} and the full covariance matrix Σ_{im} for Gaussian component m in hidden state i, where $m = 1, \ldots, M$, $i = 1, \ldots, N$.

The model parameters are adjusted in such a way that the likelihood $P(O_c|\lambda_c)$ is maximized given training data set O_c, which denotes the weight coefficient vector sequences along the temporal axis for action class c. The Baum-Welch algorithm [15] is used for iteratively re-estimate model parameters to achieve the local maximum. Given a test sequence for an unknown action with the corresponding landmark points trajectories, we first apply the non-rigid factorization to compute the deformation coefficients O. It should be noted that the basis shapes should keep same as obtained during training procedure. That is to say, we only need to iteratively estimate the rotation matrix and the weigh coefficients. Then we use maximum likelihood approach for the classification:

$$argmax_{c\in\{1,\ldots,C\}} P(O|\lambda_c) \tag{8}$$

4. Experiments

As in [8]–[11], this work does not address the lower-level



(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Fig. 2 The examples of the recovered weight coefficients. (a)–(c), (d)–(f) and (g)–(i) are for walking, running and jumping sequences respectively. For each action category, the left figure and middle figure are recovered weight coefficients for sequences of two different performers with stationary cameras under the same viewpoint while the right figure is the recovered weight coefficients for sequences from a moving on-the-fly camera. The horizontal axis represents the frame index. The vertical axis represents the value of the weights. Red, Green and Blue color represent l_1, l_2 and l_3 respectively. See the main text for detail discussion.

processing tasks such as robust body-joint detection and tracking. We concentrate on showing the discriminative power of the weights coefficient vectors feature for action recognition under varying viewpoint directions and different execution speed, given the 2D trajectories of anatomical landmarks on human body. Experiments were performed on CMU Mocap motion capture data of real human action sequences. The motion capture data is obtained by placing several sensories on the body-joints of the performer and the 3D trajectory positions of those marks are recorded. During the experimental procedure, only the 2D projected observations were used and the Z information was discarded. The dataset used in our experiment consists of eight representative real life action categories with each category has 10 sequences performed by different persons, which includes walking, running, dribbling, kicking, boxing, jumping, wheeling and dancing. The sample example views seen by the cameras are depicted in Fig. 3. In order to verify the proposed claim that the low dimensional weight coefficients vector sequence is discriminative for recognizing actions with changing viewpoint, for those sequences that the movement of the capturing camera are not large enough, for example in Fig. 4 (a), the 2D feature point trajectories were re-generated from 3D positions with projections using varying rotation and translation matrices. The variation limitation range of the rotation parameters is between [0° − 45°] for α_x, α_y and [0° − 90°] for α_z, which is large enough to make the 2D trajectory observations appear quite different for each action instances, for example in Fig. 4 (b). Again, the Z information is discarded and only the 2D trajectory

Table 1 The confusion matrix table and recognition rate. **A1**–**A8** represent walking, running, dribbling, kicking, boxing, jumping, wheeling and dancing, respectively.

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
A1	17	1	1	1	0	1	0	1
A2	2	17	1	0	0	0	0	0
A3	1	2	18	0	0	1	0	0
A4	0	0	0	19	0	0	0	0
A5	0	0	0	0	19	0	0	0
A6	0	0	0	0	0	18	0	0
A7	0	0	0	0	0	0	20	0
A8	0	0	0	0	1	0	0	19
Rate	**85%**	**85%**	**90%**	**95%**	**95%**	**90%**	**100%**	**95%**

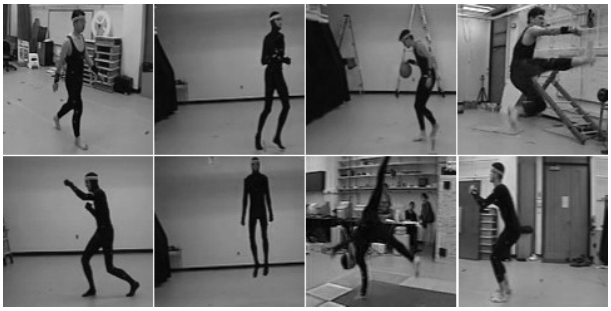
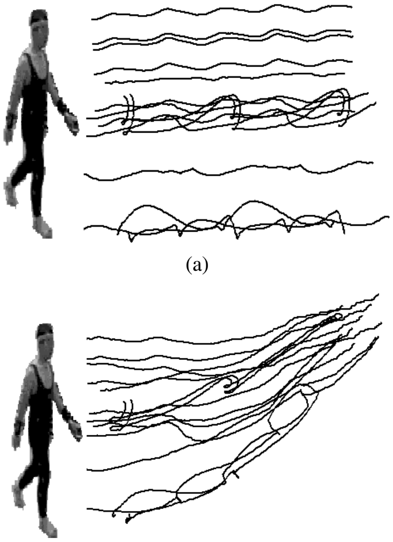


Fig. 3 The sample example views of walking, running, dribbling, kicking, boxing, jumping, wheeling and dancing.



(a)

(b)

Fig. 4 Comparison of 2D landmark points trajectories captured using static camera (a) and moving camera (b).

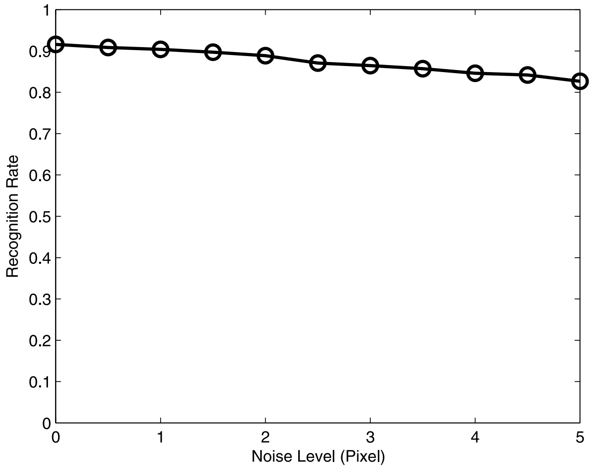


Fig. 5 Recognition rate versus coordinate noise.

observations are used as inputs of the experiment. We used the HMMs with the topology of 6 hidden states and each observation was modeled by mixtures of 3 Gaussian densities. K, which denotes the number of basis shapes, was empirically set to 3. The experiments are repeated for 10 times while in each time the sequences for each action category was randomly 80/20 partitioned into training and testing sets. Table 1 shows the results of action recognition using the proposed view invariant recognition framework. It can be clearly seen from the confusion matrix that the proposed view invariant human action recognition framework achieves promising performance and the overall recognition rate is 91.88%.

Since usually the trajectories produced by feature points tracker are not so accurate, we also tested the performance of the proposed method against coordinate noise of different levels. Specifically, we disturbed the coordinates of the feature points trajectories using uniform noises of 0–5 pixels levels with 0.5 pixel interval. Figure 5 shows the curve of recognition rate versus noise level. It can be seen that the proposed method is robust to the coordinates noise to some extent. This is expected since the extracted weight coefficients depict the motion style from a global point of view and HMMs represent the temporal changing mode in a probabilistic way.

5. Conclusion and Future Works

This paper introduces a new framework for view invariant human action recognition using 2D observations of trajectories of body landmarks, based on nonrigid matrix factorization and Hidden Markov Models. The feature vectors are extracted via non-rigid factorization by treating all of the training sequences under the same conditions. The extracted low dimensional weigh coefficients encode the discriminative information for action recognition. Based on those features, a set of HMMs are built for each action category. Promising recognition results show that the proposed algorithm is robust to noises, and more importantly, to the variations in viewing direction and execution rate.

The camera projection procedure in this paper is described using the simple orthographical model. In real life applications, the projective model is more faithful. Also, the feature point tracker often produces trajectories with missing data due to occlusion or ambiguity. How to extend the proposed method to the projective case and missing data case is worthy of further research.

References

[1] A.F. Bobick and J.W. Davis, “The recognition of human movement using temporal templates,” IEEE Trans. Pattern Anal. Mach. Intell., vol.23, no.3, pp.257–267, 2001.

[2] A. Efros, A. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” IEEE International Conference on Computer Vision, pp.726–733, 2003.

[3] T. Kobayashi and N. Otsu, “Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation,” International Conference on Pattern Recognition, vol.4, pp.741–744, 2004.

[4] A. Oikonomopoulos, I. Patras, and M. Pantic, “Kernel-based recognition of human actions using spatiotemporal salient points,” IEEE Int. Conf. on Computer Vision and Pattern Recognition, June 2006.

[5] A. Oikonomopoulos, I. Patras, and M. Pantic, “Human action recognition with spatiotemporal salient points,” IEEE Trans. Syst. Man Cybern. B, Cybern., vol.36, no.3, pp.710–719, 2006.

[6] V. Kellokumpu, M. Pietikinen, and J. Heikkil, “Human activity recognition using sequences of postures,” Proc. IAPR Conference on Machine Vision Applications (MVA 2005), pp.570–573, Tsukuba, Japan, 2005.

[7] C. Rao, A. Yilmaz, and M. Shah, “View-invariant representation and recognition of actions,” Int. J. Comput. Vis., vol.50, no.2, pp.203–226, 2002.

[8] V. Parameswaran and R. Chellappa, “View invariants for human action recognition,” Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol.2, pp.613–619, 2003.

[9] A. Gritai, Y. Sheikh, and M. Shah, “On the invariant analysis of human actions,” International Conference on Pattern Recognition, 2004.

[10] A. Yilmaz and M. Shah, “Action sketch: A novel action representation,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.984–989, 2005.

[11] A. Yilmaz and M. Shah, “Recognizing human actions in videos acquired by uncalibrated moving cameras,” Tenth IEEE International Conference on Computer Vision (ICCV’05), vol.1, pp.150–157, 2005.

[12] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: A factorization method,” Int. J. Comput. Vis., vol.9, no.2, pp.137–154, 1992.

[13] C. Bregler, A. Hertzmann, and H. Biermann, “Recovering non-rigid 3d shape from image streams,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.690–696, HiltonHead, South Carolina, June 2000.

[14] L. Torresani, D. Yang, E. Alexander, and C. Bregler, “Tracking and modeling non-rigid objects with rank constraints,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.493–500, Kauai, Hawaii, 2001.

[15] R. Lawrence and A. Rabiner, “Tutorial on hidden Markov models and selected applications in speech recognition,” Proc. IEEE, vol.77, no.2, pp.257–286, 1989.

Xi Li His primary research is on pattern recognition, computer vision and machine learning. He received his Bachelors degree in Electrical Engineering from Xi’an Jiaotong University in 1995, and his PhD in Computer Science from Xi’an Jiaotong University in 2005. From 2006 to 2008, he worked as a postdoctoral researcher in the Tsukuba University, Japan.

Kazuhiro Fukui received the B.E. and M.E. degrees in mechanical engineering from Kyushu University, Fukuoka, Japan, in 1986 and 1988, respectively. And in 2003, he received his PhD in engineering from Tokyo Institute of Technology. He worked for TOSHIBA Corporate Research and Development Center from 1988 to 2004. Since 2004 he has been an associate professor at Department of Computer Science, University of Tsukuba. His research interests include machine vision, pattern recognition and human interface technology. He is a member of the Information Processing Society Japan and the IEEE.