MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

View-invariant Human Action Recognition Based on Factorization and HMMs

Xi Li , Kazuhiro Fukui
Graduate School of Systems and Information Engineering,
University of Tsukuba, JAPAN
{xili@viplab.is,kfukui@cs}.tsukuba.ac.jp

Abstract

One of the fundamental challenges of human action recognition is accounting for the variability that arises during video capturing. For a specific action class, the 2D observations of different instances might be extremely different due to varying viewpoint when the sequences are captured by moving cameras. The situation is even worse if the actions are executed at different rates. In this paper, a novel view-invariant human action recognition method is proposed based on non-rigid factorization and Hidden Markov Models (HMMs). By assuming that the execution of an action can be approximated by dynamic linear combination of a set of basis shapes, we show that the weight coefficients of basis shapes by measurement matrix non-rigid factorization contain crucial information for action recognition regardless of the viewpoint. Based on the extracted discriminative features, the HMMs is used for action modeling and classification. The performance of the proposed method has been successfully demonstrated experimentally using real sequences.

1. Introduction

In recent years, the recognition of human gestures and actions has become an active research area within computer vision community due to its potential applications such as video surveillance, human-computer interface, content based retrieval and sports video analysis.

Many approaches for human action recognition have been presented. The most common one taken by the researcher is to perform action recognition using 2D observation. For example, A. F. Bobick et. al. used temporal templates for human movement representation and recognition[1]. A.Efros et. al. introduced the motion descriptor based on optical flow measurements in a spatio-temporal volume for each stabilized human figure, and an associated similarity measurement was used in a nearest-neighbor framework[2]. Takumi Kobayashi et al used cubic higher-order local auto-correlation for action and simultaneous multiple-person identification[3]. V. Kellokumpu et. al. presented a real-time system for recognition of 15 different continuous human activities[4]. Both of the above methods are viewpoint dependent. The training sequences and testing sequences are captured under the same viewing direction by stationary cameras. But in real life applications, for a specific action class, the 2D observations of different instances might be extremely different due to varying viewpoint when the sequences are captured by moving cameras. The situation is even worse if the actions are executed at different rates.

Some of the view-invariant methods had been proposed. C. Rao et. al. presented a computational representation of human action using spatio-temporal curvature of 2-D trajectory[5]. V. Parameswaran et. al. proposed a 3D based approach for view-invariant human action recognition[6]. A. Gritai et al used the epipolar geometric constraints computed from the correspondences of human body landmarks to match actions performed from different viewpoints and in different environments[7]. A. Yilmaz et. al. represented the action by a set of descriptor computed from a spatio-temporal action volume created from a set of object silhouette[8]. Again, the epipolar geometry between the views of two stationary cameras is exploited to achieve view-invariant recognition. The above view-invariant action recognition methods have the limitation that action sequences are captured using stationary cameras. A. Yilmaz et. al. further extended the standard epipolar geometry to the geometry of dynamic scenes where the cameras are moving[9].

In this paper, a novel view-invariant human action recognition method is proposed based on non-rigid factorization and Hidden Markov Models (HMMs). By assuming that the execution of an action can be approximated by dynamic linear combination of a set of basis shapes, we show that the weight coefficients of basis shapes by measurement matrix non-rigid factorization contain crucial information for action recognition regardless of the viewpoint. Based on the extracted discriminative features, the HMMs，which allows for the inclusion of dynamics, is used for action modeling and classification. The performance of the proposed method has been successfully demonstrated experimentally using real sequences.

This paper is organized as follows: Section 2 describes the 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 dataset are presented in section 4, followed by conclusions in section 5.

2. Feature extraction based on non-rigid factorization

As in [6,7,8,9], 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 of view-invariant is not obvious. It is difficult to recognize actions captured by moving 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. Fig. 1 shows an example using sample walking sequences. Fig. 1(a) and 1(b) are two walking sequences performed by different person. Fig. 1(c) and 1(d) are the 2D trajectories observations for the two walking sequences under same viewing directions by stationary camera, respectively. It can be seen that even the two sequences are performed by different persons, the 2D observations still look similar since they belong to the same action class and the body joints move in a consistent way. Fig. 1(e) and 1(f) are the 2D trajectories observations for the two walking sequences projected using moving cameras, with the trajectories superimposed. Due to the motion of the camera, it is evident that not only the trajectories in Fig.1(e) and 1(f) do not appear similar, but also the trajectories in Fig.1(c) and 1(e) look quite different even these sequences performed by a same person belong to the same action category.

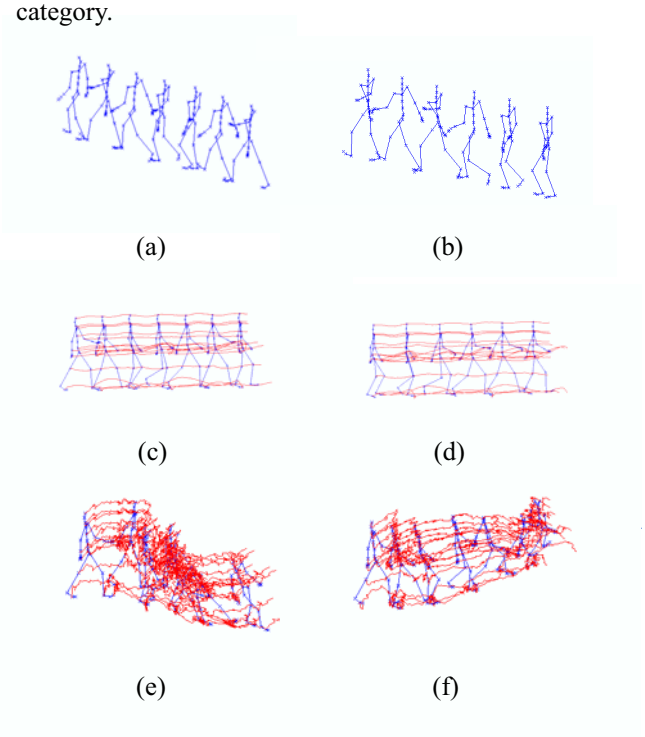

Figure 1: Example of sample walking sequences. (a)(b): 3D sequences; (c)(d): 2D trajectories with stationary cameras; (e)(f): 2D trajectories with moving cameras; for two different performers respectively

Our approach for recognizing human actions in videos acquired by moving cameras is based on the observation that a deformable shape(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 deformation coefficients of basis shapes contain the crucial 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[10]. The problem in the action recognition scenario is more complex because the freedom for moving human body is extremely high due to the non-rigidity. C. Bregler[11] and L. Torresani[12] further extended the factorization method for non-rigid case. Suppose that P feature landmarks are tracked across F frames, the deforming shape can be described as a key frame basis set $S_1, S_2, ..., S_K$, 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:

$$S = \sum_{i=1}^{K} l_i S_i \qquad S, S_i \in R^{3XP}, l_i \in R \qquad (1)$$

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

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

(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 is the camera translation. It can be rewritten as follows after eliminating T by subtracting the mean of feature points as in [10]:

$$\begin{bmatrix} u_1 & u_2 & \cdots & u_P \\ v_1 & v_2 & \cdots & v_P \end{bmatrix} = \begin{bmatrix} l_1 R & l_1 R & \cdots & l_K R \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_K \end{bmatrix} \qquad (3)$$

If we write all the feature points along the temporal axis into a 2FxP matrix W:

$$W = \begin{bmatrix} u_1^1 & u_2^1 & \cdots & u_P^1 \\ v_1^1 & v_2^1 & \cdots & v_P^1 \\ u_1^2 & u_2^2 & \cdots & u_P^2 \\ v_1^2 & v_2^2 & \cdots & v_P^2 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ u_1^F & u_2^F & \cdots & u_P^F \\ v_1^F & v_2^F & \cdots & v_P^F \end{bmatrix} \qquad (4)$$

W is the measurement matrix and can be further decomposed into the following form:

$$W = \begin{bmatrix} l_1^1 R^1 & l_2^1 R^1 & \cdots & l_K^1 R^1 \\ l_1^2 R^2 & l_2^2 R^2 & \cdots & l_K^2 R^2 \\ \vdots & \vdots & \vdots & \vdots \\ l_1^F R^F & l_2^F R^F & \cdots & l_K^F R^F \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_K \end{bmatrix} \qquad (5)$$

L. Torresani[12] proposed an effective way for fac-

torization of the measurement matrix W as above equation. First, the weighting coefficients $l_k^t, k=1,...,K, t=1,...,F$ are randomly initialized, and then the shape bases $S_i, i=1,...K$ 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[12].

Denote the weight coefficient vector corresponding to frame i as $L^{(i)} = (l_1^i, l_2^i, ..., l_K^i)$, then the vector sequence $\phi = (L^{(1)}, L^{(2)}, ..., L^{(F)})$ contains the necessary information for action recognition regarding the human body movement. ϕ describes the changing mode for the body-parts. The ϕs for different action categories should exhibit different patterns while the ϕs for same action should have similar patterns. But 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 framework, 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:

$$W = [W_1^{1T}, \cdots, W_1^{N_1 T}, \cdots, W_C^{1T}, \cdots, W_C^{N_C T}]^T \qquad (6)$$

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 $\phi_i^j, i=1,...,C, j=1,...,N_i$. 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,

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

Since the different actions share the same shape basis, the discriminative information for action recognition are encoded in the ϕ_i^j s. Fig. 2 shows the examples of the recovered weight coefficients for different action classes. Fig. 2(a) and (c) are for walking sequences of two different performers with stationary viewing camera, while Figure2(e) is also for walking sequence but projected from a moving camera. Figure2(b) (d) and (f) are for running case under the same conditions like the walking case. It can be seen that varying patterns of the weight coefficients varying curves 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 cameras.

3. Action Modeling and Recognition using HMMs

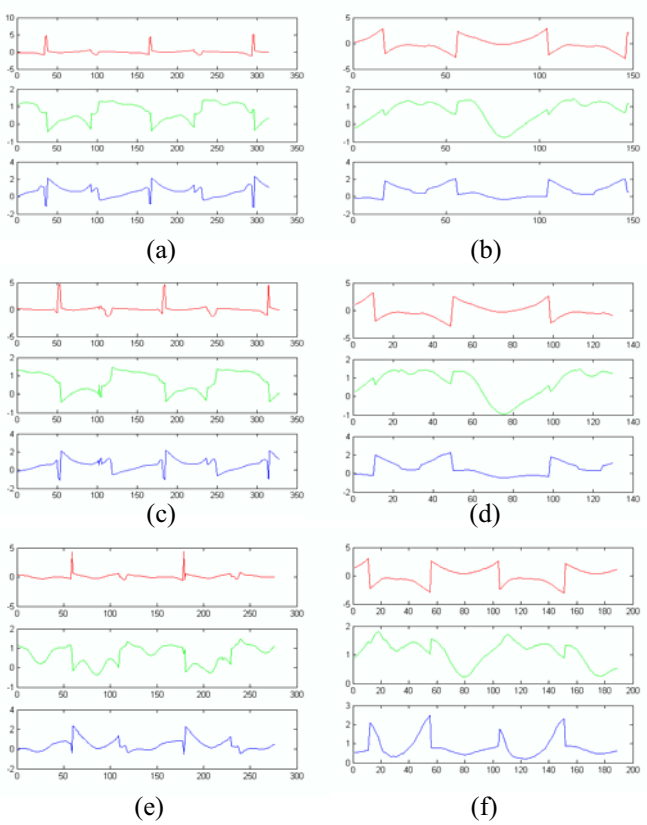

Figure 2: The examples of the recovered weight coefficients. (a)(c) are for walking sequences by two different performers with stationary viewing camera. (e) is also for walking sequence but projected from a moving camera. (b)(d)(f) are for running case under the same conditions like the walking case.

Hidden Markov Models(HMMs) have been successfully used for speech recognition and computer vision[13]. We employ the HMMs for action modeling and recognition because 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 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 $m = 1,...M, i = 1,...,N$.

The model parameters are adjusted in such a way that the likelihood $P(O_c \mid \lambda_c)$ is maximized by using the given set of training data set O_c, which denotes the weight coefficient vector sequences along the temporal axis for action class c. The Baum-Welch algorithm[13] is used for iteratively re-estimate model parameters to achieve the local maximum.

Given a test sequence for an unknown action with the corresponding feature vector sequence O, we first apply the non-rigid factorization to compute the deformation coefficients. It should be noted that the basis shape 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:

$$\arg\max_{c \in \{1,...,C\}} P(O \mid \lambda_c) \qquad (8)$$

4. Experiments

Experiments are performed on CMU human motion capture data for real action sequences. It should be noted that during the experiments procedure, only the 2D projected observations are used and we did not use any Z information. The dataset used in our experiment consists of eight representative classes of actions with each class has several sequences performed by different persons, which are evenly split into training sets and testing sets. The eight action classes include "walking", "running", "dribbling", "kicking", "boxing", "jumping", "wheeling" and "dancing". For the purpose of verifying the claim in this paper that the weight coefficients vector sequence is discriminative for recognizing actions in varying viewpoint, the 2D feature point trajectories are computed with projections using randomly generated rotation matrixes. The purpose is to simulate the real life conditions of recognizing actions using image sequences captured by moving cameras. We use the HMMs with the topology of 6 hidden states and each observation is modeled by using mixtures of 3 Gaussian densities. K, which denotes the number of basis shapes, is empirically set to 3. In table 1 we give the results of action recognition using the proposed view-invariant recognition framework. The experiments are repeated for 20 times while in each time the whole dataset are randomly split into training and testing sets. It can be clearly seen that the proposed method works robustly under the condition that the capturing cameras are moving.

Table 1: Recognition rate

Action	walking	running	dribbling	kicking
Rate	93.75	92.86	95.00	96.67
Action	boxing	jumping	wheeling	dancing
Rate	95.50	95.00	95.00	96.25

5. Conclusion

In this paper, we propose a novel method for HMMs-based view-invariant human action recognition. The feature vectors are extracted via non-rigid factorization by treating all of the training sequences under the same ground. The extracted weigh coefficients encode the discriminative information for action recognition. Based on those features, a set of HMMs were built for each action category. The recognition results are convincing and show that our algorithm is robust to the variations in viewing direction and execution rate.

References:

[1] A. F. Bobick and J. W. Davis, The Recognition of Human Movement Using Temporal Templates, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, 2001.

[2] A.Efros, A.Berg,G.mori, and J.Malik, Recognizing action at a distance. In IEEE International Conference on Computer Vision, pp. 726-733, 2003

[3] Takumi Kobayashi, Nobuyuki Otsu, Action and Simultaneous Multiple-Person Identification Using Cubic Higher-Order Local Auto-Correlation, 17th International Conference on Pattern Recognition (ICPR'04) - Volume 4, pp. 741-744, 2004.

[4] V. Kellokumpu, M. Pietikäinen and J. Heikkilä, Human activity recognition using sequences of postures, Proc. IAPR Conference on Machine Vision Applications (MVA 2005), Tsukuba Science City, Japan, pp.570-573., 2005

[5] C. Rao, A. Yilmaz, M. Shah, View-Invariant Representation And Recognition of Actions, International Journal of Computer Vision, Vol. 50, Issue 2, pp . 203-226,2002

[6] V. Parameswaran and R. Chellappa, View Invariants for Human Action Recognition, Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Madison, WI, Vol. 2, pp. 613-619, June 2003.

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

[8] A. Yilmaz and M. Shah, Action sketch: A novel action representation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 984-989, 2005.

[9] A. Yilmaz, Mubarak Shah, Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras, Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, pp. 150-157, 2005.

[10] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: A factorization method, International Journal of Computer Vision, vol.9, no. 2, pp.137-154, 1992.

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

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

[13] R. Lawrence, and A. Rabiner, Tutorial on hidden markov models and selected applications in speech recognition, Proc. of the IEEE, vol. 77, no.2.,pp.257-286, 1989