

# Full Body Tracking-Based Human Action Recognition

Gu Junxia Ding Xiaoqing Wang Shengjin Wu Youshou

State Key Laboratory of Intelligent Technology and Systems

Tsinghua National Laboratory for Information Science and Technology

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

{gujx, dxq, wsj, wuys}@ocrserv.ee.tsinghua.edu.cn

## Abstract

*In this paper, we present a novel method for human action recognition with the combined global movement feature and local configuration feature. The human action is represented as a sequence of joints in the 4D spatio-temporal space, and modeled by two HMMs, a conventional HMM for global movement feature and an exemplar-based HMM for configuration feature. Firstly, an adaptive particle filter is adopted to track the marker-less actor's 3D joints. Then, the combined features are extracted from the full body tracking results. Finally, the actions are classified by fusing two HMMs. The effectiveness of the proposed algorithm is demonstrated with experiments on 7 actions by 12 actors. The results prove robustness of the proposed method with respect to viewpoints and actors.*

## 1. Introduction

Human action recognition is receiving increasing attention in the past decade. However, the various orientations of actors and complex actions make the processing very challenging.

Depending on the action representation, the existing algorithms can be classified into two types, template-based and model-based. Template-based approaches [1, 2, 3] directly represent actions using image information, such as silhouettes or optical flow. This method always limits recognition to the situations where observed and learned processes are obtained using similar camera configurations [4]. In contrast, model-based approaches [5, 6, 7] assume a known human model, and represent actions in a joint space. Johansson demonstrated that a simple point-based model of the human body contained sufficient information for the recognition of actions [8]. Recently

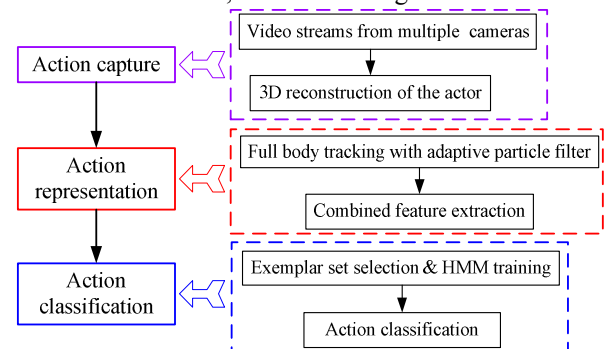
the improvement of the marker-less full body tracking makes these approaches available.

In this paper, we propose a model-based approach. The action is modeled as a sequence of joints in the 4D spatio-temporal space. Firstly, full body tracking approach is used to obtain the actor's global movement information and local configuration feature. Then, we combine a conventional HMM for movement feature and an exemplar-based HMM for configuration feature to classify the actions.

The paper proceeds as follows. Section 2 presents an overview of the proposed approach. Section 3 describes the action representation based on full body tracking. Section 4 focuses on the action recognition method. Section 5 shows the experimental results. Section 6 concludes this paper.

## 2. Action recognition system

The action recognition system includes three basic problems: action capture, action representation, and action classification, as shown in Figure 1.



**Figure 1. Chart of action recognition system**

**Action capture.** Multiple video streams are simultaneously captured from static calibrated cameras. And foreground/background segmentation is

performed on each through background subtraction method. Then, volumetric representation sequences of the 3D actor to be tracked are created.

**Action representation.** 3D joints sequence is one of the most effective methods of action representation. Full body tracking algorithm is adopted to track actor's joints. However, the human body's large degrees of freedom (DOF) and the nonlinearity of the dynamic system bring many difficult problems. To overcome these limitations, we propose a tracking approach which fuses the body part segmentation and adaptive particle filter. Then we extract the movement feature (actor's location and orientation) and configuration feature (normalized actor's joints).

**Action Classification.** We represent each action using two HMMs. One is a conventional HMM for movement feature and the other is an exemplar-based HMM [4] for configuration feature. The key poses included in the action sequences provide the foundation of the exemplar-based HMM.

### 3. Action representation

An articulated human model, which consists of approximately 4500 vertices, is adopted in the tracking method. Its skeletal structure is modeled by 19 DOF, as shown in Figure 2. In addition, there is 1 DOF for the actor's orientation (rotation angle around Z axis), and 3 DOF for the actor's location.

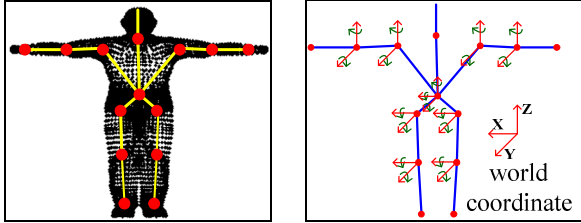


Figure 2. Human model and its DOF

Figure 3 illustrates the full body tracking algorithm. The body part segmentation and body tracking are coupled. Body part segmentation uses the tracking result of the previous frame, and each part's tracking relies on the segmentation result.

#### 3.1. Adaptive particle filter for body tracking

Semi-supervised clustering method is used to segment the actor's body parts as shown in Figure 3. Literature [9] presented this approach in detail. With the segmentation result, particle filter is adopted to track each body part.

In the action sequences, the movement of each body part is often different. For example, in the 'wave' action, the arm's movement is large, while the leg's

movement is small. Then adaptive particle number should be arranged to each body part.

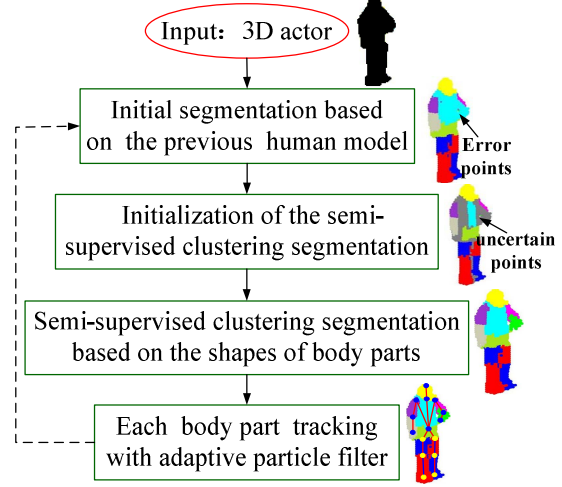


Figure 3. Full body tracking algorithm

**3.1.1. Adaptive particle number.** The particle number always varies with the noise variance and the survival rate. For particle set  $(x_i, \omega_i)_{i=1}^N$ , survival rate [10] is defined as

$$\mu = \frac{N_{eff}}{N} \quad (1)$$

where  $N_{eff} = 1 / \sum_{i=1}^N (\omega_i)^2$ . Then the particle number  $N$  needed can be obtained from:

$$N = \frac{N_{min}}{\mu} \quad (2)$$

where  $N_{min}$ , which is directly proportional to the variance, represents the minimum acceptable effective particle number for successful tracking of an object.

We embed the adaptive particle number into the traditional particle filter approach [11].

**3.1.2 Probabilistic state inference.** Define the particle's state  $x$  as the rotation angle vector and  $(x_i, \omega_i)_{i=1}^N$  as the particle set.  $Z_k$ , the points set of a body part in 3D actor, is the measurement at time  $k$  and  $T(x_i)$  is the corresponding body part in human model with the pose of  $x_i$ .  $d(Z_k, T(x_i))$  indicates the distance between  $Z_k$  and  $T(x_i)$ . In prediction stage, the state is propagated through the first order dynamic model. And we use the sigmoid function to approximate measurement likelihood:

$$p(Z_k | x_i) = \frac{1}{1 + \exp(d(Z_k, T(x_i)))} \quad (3)$$

### 3.2. Feature extraction

At time  $k$ , we define  $g(k) = (l(k), o(k))$  as the movement feature, where  $l(k)$ ,  $o(k)$  respectively denote the actor's location and orientation, and  $c(k) = (j_i(k))_{1 \leq i \leq M}$  as the configuration feature, where  $j_i(k)$  is the normalized 3D coordinate of the actor's  $i_{th}$  joint.

Based on the full body tracking result, we can extract these features easily.

### 4. Action recognition

#### 4.1 HMM Learning

A conventional HMM for movement feature is trained with the standard Baum-Welch algorithm [12]. And for the exemplar-based HMM, the exemplar selection and the HMM learning are performed alternately [4]. Let  $X$  be the exemplar set, and  $\zeta$  denotes the training set of all the actions. The detail algorithm is shown as follows:

**Step1:** Set  $X = \emptyset$

**Step2:** Find  $y^* \in \{\zeta \setminus X\}$ , where a classifier using exemplar set  $\{X \cup y^*\}$  has best recognition performance in the training set,  $\zeta$ .

**Step3:** Repeat step 2 until  $K$  exemplars selected.

In the exemplar-based HMM learning, each model is learned through the Baum-Welch algorithm for Gaussian mixture HMM [12], except that the means of the Gaussians ( $X$ ), are not updated.

#### 4.2 Action recognition

We have learned two action models, conventional HMM  $\lambda_c^1$  and exemplar-based HMM  $\lambda_c^2$ , for each action class  $c \in \{1, \dots, C\}$ . An action sequence  $Y = \{y_1, \dots, y_T\}$  is then classified with MAP estimate:

$$g(Y) = \arg \max_c (\alpha_c \times p(\lambda_c^1 | Y) + (1 - \alpha_c) \times p(\lambda_c^2 | Y)) \quad (4)$$

where  $\alpha_c$  is a weight for the action class  $c$ .

**Table 1. Recognition rates**

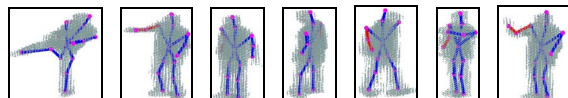
feature	kick	point	stand	turn	walk	watch	wave	average
movement feature	35%	31%	24%	99%	99%	72%	6%	52%
configuration feature	100%	100%	67%	100%	100%	100%	100%	95%
combined features	100%	100%	100%	100%	100%	100%	100%	100%

Figure 6 shows the confusion matrix. Among 7 actions to be classified, the most confusion occurs

### 5. Experimental results and analysis

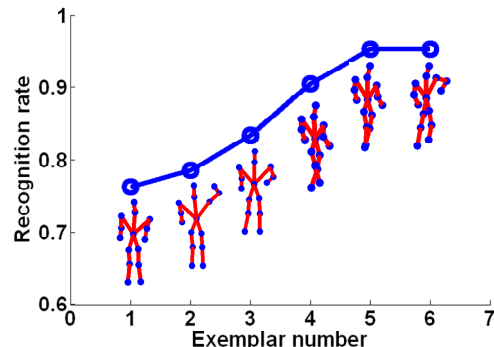
This paper presents results on an action dataset from IXMAS dataset [13]. We choose 7 actions, performed by 12 actors, each 3 times. The volume of interest is divided into  $64 \times 64 \times 64$  voxels.

Figure 4 shows the 3D actor's full body tracking results. The tracker can successfully obtain the joints under various orientations. The average error of the joints is 0.70 voxel. On standard PC (Pentium IV 3.0GHz CPU), the proposed method runs at about 4.4s/f.



**Figure 4. 3D actors and their joints (Form left to right, the actions are 'kick', 'point', 'stand', 'turn around', 'walk in circle', 'check watch', and 'wave').**

Ten of the actors are used for model learning and exemplars selection, and the remaining two are used for testing. The number of exemplars is empirically set to 6, as shown in Figure 5. Actions are modeled with 3 states.



**Figure 5. Selected exemplars and recognition rate in the testing set vs. number of exemplars.**

Table 1 shows the recognition rates, where we use movement feature ( $\alpha_c = 1, 1 \leq c \leq 7$ ), configuration feature ( $\alpha_c = 0, 1 \leq c \leq 7$ ), and combined features ( $\{\alpha_c\}_{c=1:7} = \{0.2, 0.2, 0.2, 0.5, 0.5, 0.2, 0.2\}$ ). Average recognition rates with the three kinds of features are 52%, 95%, and 100%, respectively.

between 'stand' and 'turn around' with the configuration feature. With the movement feature, the

'stand' and 'turn around' actions can be separated easily. However the other actions are confused. With both kinds of features, these actions can be classified effectively.

We compare our human action recognition approach with other previous researches as given in

(a)	kick	point	stand	turn	walk	watch	wave	(b)	kick	point	stand	turn	walk	watch	wave	(c)	kick	point	stand	turn	walk	watch	wave
kick	0.35	0.21	0.06	0.11	0.01	0.07	0.19	kick	1	0	0	0	0	0	0	kick	1	0	0	0	0	0	0
point	0.19	0.31	0.13	0.01	0	0.19	0.17	point	0	1	0	0	0	0	0	point	0	1	0	0	0	0	0
stand	0.01	0.03	0.24	0	0	0.66	0.06	stand	0	0	0.67	0.33	0	0	0	stand	0	0	1	0	0	0	0
turn	0	0	0	0.99	0.01	0	0	turn	0	0	0	1	0	0	0	turn	0	0	0	1	0	0	0
walk	0	0	0	0.01	0.99	0	0	walk	0	0	0	0	1	0	0	walk	0	0	0	0	1	0	0
watch	0.05	0	0.14	0	0	0.72	0.09	watch	0	0	0	0	0	1	0	watch	0	0	0	0	0	1	0
wave	0.13	0.04	0.17	0	0	0.60	0.06	wave	0	0	0	0	0	0	1	wave	0	0	0	0	0	0	1

Figure 6. Confusion matrices. (a) movement feature; (b) configuration feature; (c) Combined features

method	action	actor	feature	view	Recognition rate
Davis et al. [1]	18	1	MEI & MHI template	multi-view	83.3%
Wang et al. [2]	10	9	MMS & AME template	single view	96.7%
Ahmad et al. [3]	7	11	optic flow & shape flow	multi-view	88.29%
Weinland et al. [4]	11	10	silhouette & 3D visual hull	multi-view	81.3%
Ren et al. [5]	7	—	3D motion	multi-view	about 90%
Our approach	7	12	3D joints	multi-view (3D)	100%

Figure 7. Comparison of the proposed approach with some previous researches

## 6. Conclusion

This paper addressed a robust human action recognition approach based on the full body tracking result by using combined global movement feature and local configuration feature. Firstly, an adaptive particle filter is adopted to track the full body. Then the combined features are extracted from the result of the full body tracking. Corresponding to the two kinds of features, a conventional HMM and an exemplar-based HMM are built for each kind of actions. This approach is evaluated on a dataset of 12 actors, and 7 actions. The combined features are able to effectively classify these mentioned actions. The applications of action recognition require the development of systems which are real-time and robust to environment variation. However, the feature extraction through full body tracking is time-consuming. Our future work includes the acceleration of the full body tracking approach and the analysis of the complex multiple human actions.

## References

[1] J. W. Davis, and A. F. Bobick. The representation and recognition of human movement using temporal templates. *CVPR*, pp: 928 - 934, June 1997.

[2] Liang Wang and David Sute. Informative Shape Representations for Human Action Recognition. *ICPR*, pp: 1266 - 1269, August 2006.

[3] Mohiuddin Ahmad, Seong-Whan Lee. Human action recognition using shape and CLG-motion flow from

Figure 7. It is difficult to compare these approaches, since the data sets and environments are different. However, the results can give a general overview and comparison of some approaches in action recognition. Compared to the mentioned researches, our approach yields better results.

multi-view image sequences. *Pattern Recognition*. 41(7): 2237-2252, July 2008.

[4] Daniel Weinland, Edmond Boyer, Remi Ronfard. Action Recognition from Arbitrary Views using 3D Exemplars. *ICCV*, pp: 1-7, Oct. 2007.

[5] Haibing Ren, Guangyou Xu. Human Action Recognition with Primitive-based Coupled-HMM. *ICPR*, pp: 494 - 498, August 2002.

[6] Alexei Gritai, Yaser Sheikh and Mubarak Shah. On the use of Anthropometry in the Invariant Analysis of Human Actions, *ICPR*, pp:923-926, August 2004.

[7] Alper Yilmaz, Mubarak Shah. Matching actions in presence of camera motion. *Computer vision and image understanding*, 104(2-3): 221-231, 2006.

[8] G. Johansson. Visual motion perception. *Scientific American*, 232(2): 76-88, 1975.

[9] Gu Junxia, Ding Xiaoqing, and Wang Shengjin. A Semi-supervised clustering-based segmentation algorithm of 3D reconstructed human body parts. *Journal of Image and Graphics*. 13(3): 558-565, 2008. (in Chinese)

[10] John MacCormick and Michael Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. *ECCV*, pp: 3-19, June 2000.

[11] M.Sanjeev Arulampalam, Simon Maskell, Neil Gordon, et al. A tutorial on particle filter for online nonlinear / non-Gaussian Bayesian tracking. *IEEE Trans. on Signal Processing*, 50(2): 174-188, 2002.

[12] Lawrence R. Rabiner. A tutorial on hidden Markov model and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): 257-286, 1989.

[13] The IXMAS database. <https://charibdis.inrialpes.fr>.