3D-Posture Recognition using Joint Angle Representation

Adnan AL ALWANI[1] Youssef CHAHIR[1] Djamal E. GOUMIDI[2] Michèle MOLINA[3]
François JOUEN[4]

[1]GREYC CNRS (UMR 6072)
[3]PALM, EA 4649
Université de Caen, Basse-Normandie
Caen, France
[4]CHArt EA 4004
Ecole Pratique des Hauts Etudes
[13]{first name. second name@unicaen.fr}
2{gmdjml@yahoo.com}
4{francois.jouen@ephe.sorbonne.fr}

Abstract. This paper presents an approach for action recognition performed by human using the joint angles from skeleton information. Unlike classical approaches that focus on the body silhouette, our approach uses body joint angles estimated directly from time-series skeleton sequences captured by depth sensor. In this context, 3D joint locations of skeletal data are initially processed. Furthermore, the 3D locations computed from the sequences of actions are described as the angles features. In order to generate prototypes of actions poses, joint features are quantized into posture visual words. The temporal transitions of the visual words are encoded as symbols for a Hidden Markov Model (HMM). Each action is trained through the HMM using the visual words symbols, following, all the trained HMM are used for action recognition.

Keywords: Key words: 3D-joint locations, Action recognition, Hidden Markov Model, Skeleton angle

1 Introduction

Action recognition from video is considered as one of the most active research area in the field of computer vision, especially in the field of video analysis, surveillance system, and human-computer interaction. There is rich literature in action recognition in a wide range of applications, including computer vision, machine learning, and pattern recognition [22,23]. In the past years, efforts have focused on recognizing actions from video sequences with single camera. Among the different approaches, spatial-temporal interest points (STIP) and 2D binary silhouettes are the most popular representations of the human activity and action [8,14,17,18]. In the past decade, several silhouette-based methods for action recognition were mainly categorized into two subsets. One is designed to handle the sequences of silhouettes in order to extract action descriptors. Then conventional classification strategies are frequently used for

recognition [1,2,3,4]. The other category models the dynamics of the action explicitly based on the features extracted from each silhouette [5,6,7,16].

However, particular challenges in the human action recognition can alter the performance of actions descriptor from 2D image sequences: intra-class variation, inter-class dependence of action, different contexts of the same action and occlusions are the major challenges in action recognition. The use of several cameras significantly alleviates the challenges such as, occlusion, cluttered background, and viewpoints changes, which are the major low-level difficulties that reduce the recognition performance from traditional 2D imagery. Furthermore, using multiple cameras provided stable information of actions from certain viewpoints. For example, taking account a direction of the camera makes possible to distinguish object pointing from reaching from depth map rather than in RGB space. However, earlier range sensors were either difficult to use on human subjects, or may provide poor measurement. To overcome the limitations of range sensors, depth has to be inferred from stereoscopic using low-cost visible light cameras. Furthermore, 3D body configurations can be captured by multiple cameras in a predefined environment [25].

Skeleton is an articulated system, which consists of limbs segments and the joints between segments. Joints connect rigid segments and articulated motion can be considered as a continuous evolution of local poses configuration [10]. Therefore, for a given sequences of 3D maps, if we can get the stream of the 3D joints location, then reliably action recognition can be achieved by using the tracked joints locations, which significantly improves human action recognition that is under-recognized by traditional techniques.

Recently 3D information has been interpreted using special release of the Microsoft Kinect®, which provides both depth and RGB image streams. Although mainly targeted for commercial purpose, this device has brought considerable interest to the research in computer vision, and hand gesture control.

In this article, we recommend a method for posture-based human action recognition. In the proposed work, 3D locations of joints from skeleton configurations are considered as inputs. Skeletal joints positions are extracted and simple relation between coordinates vector is used to describe the 3D human poses. We perform first the representation of human postures by selecting 7 primitive joints positions. The collection of joint-angle features is quantized through unsupervised clustering into k pose vocabularies. Then encoding temporal joint-angle features into discrete symbols is performed to generate Hidden Markov Model HMM (HMM) for each action. We recognize individual human action using generated HMM. The proposed method is evaluated with public 3D dataset.

The contribution parts in this work consist of two parts: First, we use joint-angle positions to describe posture representation as human action recognition system. Second, our method presents low computational cost since only 7 joints are adopted, and includes representation of poses that is view-invariant.

The organization of this paper is as follows. We introduce the related works in section 2. Section3 describes the method we used to elaborate the architecture of proposed system from postures representation to features extraction. Section 4 addresses

action recognition by an HMM. Section 5 explains experimental results. Section 6 concludes the paper.

2 Related work

Efforts have been reported for the problem of human action recognition, by exploring different kind of visual information. Review on the categories of visual features can be found in [8,22,25]. However, only few attempts on action recognition using depth maps have been recently proposed. Therefore, we present a review of works based on 3D poses action recognition since they are related to our work.

The recent trends in the field of action recognition that use depth maps have induced further progress. Uddin et al. [13] reported a novel method of action recognition using body joint angles estimated from a pair of stereo images from stereo cameras. Thang et al. [21] developed a method for estimating body joint angles from time-series of paired stereo images recreded with a single stereo camera. Yu and Aggarwal [11] adopted an approach for action recognition where body parts are considered as a semantic representation of postures. Weinland et al. [26] proposed a model action involving 3D sequences of prototypes, which are represented as visual structures captured by a system of 5 cameras. The work proposed by Li et al.[9] suggested to map the dynamic actions as a graph, and sample a set of 3D points from the depth maps to describe a set of salient postures, that correspond to the nodes in the graph. However, the challenge in the sampling technique is view dependent. Xai et al. [12] presented a method of action recognition based on 3D skeleton joints location. They proposed a compact representation of postures by characterizing human poses as histogram of 3D joints locations sampled inside spherical coordinates system.

3 Body Parts Representation

In this section we describe the human poses representation and joints position estimation from skeleton model. This kind of representation involves 3D joints coordinates to describe a basic body structure reduced to 20 skeletal joints. Recent release of Kinect® system offers better solution for the estimation of the 3D joint positions. Figure 1 demonstrates the result of the application of depth map and the 3D skeletal joints extraction according to algorithm of Shotton et al [24] who proposed to extract 3D body joint locations from a depth map.

This algorithm is used to estimate pose locations of skeletal joints. Starting with a set of 20 joints coordinates in a 3D space, we compute a set of features to form the representation of postures. Among the 20 joints, 7 primitive joints coordinates are selected to describe geometrical relations between body parts. The category of primitive joints offers redundancy reduction to the resulting representation. Most importantly, primitive joints achieve view invariance to the resulting pose representation, by aligning the cartesian coordinates with the reference direction of the person. Moreover, we propose an efficient and view-invariant representation of postures using 7 skeletal joints, including L/R hand, L/R feet, L/R hip, and hip center.

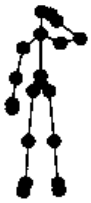

Fig. 1. (a) Depth map image. (b) Skeletal joints positions proposed by [15]

The hip center is considered as the center of coordinate system, and the horizontal direction is defined according to left hip and right hip junction. The remaining 4 skeletal locations are used for poses joint angles descriptor.

3.1 Action Coordinates Description for Skeletal Joints

View invariance is a challenging problem in action recognition. With the use of 3D body skeleton, we can capture the 3D positions of the human body. We propose a viewpoint-invariant representation of body poses by using 3D joint angles from skeletal data. In our approach of poses features inference, we achieve the view-invariant by aligning the Kinect® cartesian system with the direction of human body as shown in the Fig 2. We consider the hip center joint as the center of the new orthogonal coordinates. We define the horizontal offset vector γ to represent the vector from left to right of the hip center, the reference vertical vector ρ as the vector that is perpendicular to the horizontal reference vector computed by rotating the vector γ by 90°. The depth reference vector β is obtained by cross product operation between γ and ρ. The next steps demonstrate the procedure of aligning the orthogonal coordinates with the specific reference direction of the body.

Let the system Landmark be defined as *Rs (O,i,j,k),* and the actions landmark as *Ra(Ó, γ, ρ, β)*. If we define the hip center as origin of the action coordinates, then the action horizontal direction γ is written as:

$$\vec{\gamma} = \begin{pmatrix} hipcenter_x & + & \lambda_x \\ hipcenter_y & + & \lambda_y \\ hipcenter_z & + & \lambda_z \end{pmatrix} = \begin{pmatrix} \gamma_x \\ \gamma_y \\ \gamma_z \end{pmatrix}. \quad (1)$$

Where $\lambda = \frac{\vec{u}}{|\vec{u}|}$, is the normal unit vector, and u is defined as:

$$\overrightarrow{\boldsymbol{u}} = \begin{pmatrix} \boldsymbol{lh_x} & - & \boldsymbol{rh_x} \\ \boldsymbol{lh_y} & - & \boldsymbol{rh_y} \\ \boldsymbol{lh_z} & - & \boldsymbol{rh_z} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u_x} \\ \boldsymbol{u_y} \\ \boldsymbol{u_z} \end{pmatrix}. \qquad (2)$$

where l_h , r_h are left hip and right hip.

By performing some vector manipulations, the reference vector ρ is defined as the vertical vector that is perpendicular to the horizontal plane, and the vector β is calculated from the cross product operation between γ and ρ vectors.

For the point in the 3D coordinate system M(x,y,z), the unit vector translation from OM to ÓM is defined as:

$$\begin{aligned} \overrightarrow{\acute{O}M} &= \overrightarrow{\acute{O}O} + \overrightarrow{OM} \\ &= \overrightarrow{OM} + \overrightarrow{\acute{O}O} \\ &= M_x\vec{\imath} + M_y\vec{\jmath} + M_z\vec{k} - hc_x\vec{\imath} - hc_y\vec{\jmath} - hc_z\vec{k}. \end{aligned} \qquad (3)$$

where $h_{c,}$ is the hip center, and i,j,k are the unit direction vectors of coordinates system.

In order to express the ÓM as a function of skeletal landmarks, we first specify the system unit vectors i,j,k in terms of the action system coordinates as:

$$\begin{aligned} \vec{\imath} &= i_\gamma\vec{\gamma} + \quad j_\rho\vec{\rho} + \quad k_\beta\vec{\beta} \\ \vec{\jmath} &= i_\gamma\vec{\gamma} + j_\rho\vec{\rho} + k_\beta\vec{\beta} \\ \vec{k} &= i_\gamma\vec{\gamma} + j_\rho\vec{\rho} + k_\beta\vec{\beta}. \end{aligned} \qquad (4)$$

Substituting eq. 4 into eq. 3, we get the final formula for the vector ÓM as:

$$\overrightarrow{OM} = M_\gamma\vec{\gamma} + M_\rho\vec{\rho} + M_\beta\overrightarrow{\beta}. \qquad (5)$$

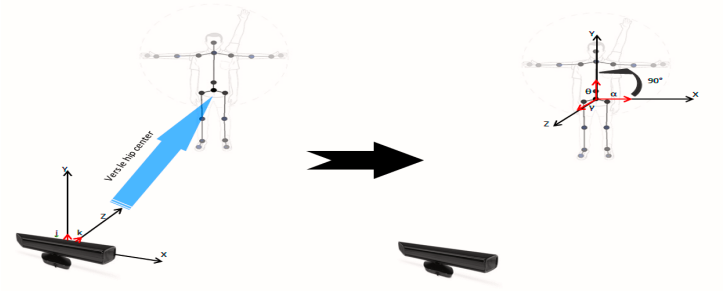

Fig. 2. Coordinates system description of skeletal joints

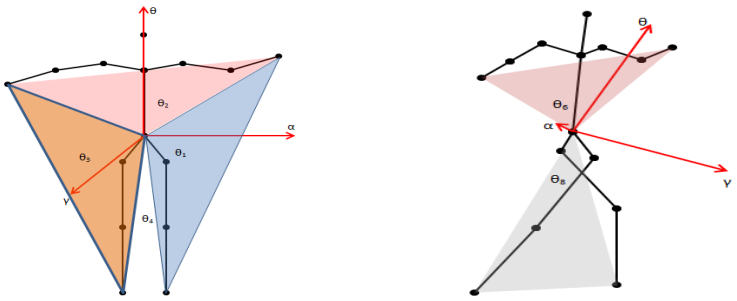

Fig. 3. Joints angles features.(a) XY plane. (b) ZY plane

3.2 Features Description

In our approach human poses are distinguished by the idea of angles groups estimated from the four junctions, which are mentioned above. The joints angles groups are sampled from two planes, XY plane and ZY plane. In the XY plane, four angles represent the angles between hand left-foot left, hand left-hand right, hand right-foot right, and foot right-foot left respectively. Same joints angles are also defined from the plane ZY. The final features vector includes eight joints angles $Ft=\{\theta_1,\theta_2,\ldots\ldots\theta_8\}$at each pose instant t. Fig. 3 shows the joint angles of two planes, where each angle is defined according to the corresponding four junctions which were illustrated in section 3.

4 HMM for Action Recognition

To apply HMMs to the problem of human action recognition, video frames $V=\{I_1,I_2,\ldots I_T\}$are transformed into symbols sequences O. The transformation is done during the learning and recognition phases. From each video frames, a feature vector $f_i \epsilon R,\{$ $(i=1,2,\ldots.T)$,T is the number of the frames$\}$ is extracted, and f_i is assigned to a symbol v_j chosen from the set of symbols V. In order to specify observation symbols, we perform the clustering of features vector into k clusters using *K*-means. Then each posture is represented as a single number of a visual word. In this way, each action is a time series of visual words. The obtained symbol sequences are used to train HMMs to learn the correct model for each action. For the recognition of a test action, the obtained observation symbol sequence $O = \{O_1,O_2,\ldots.O_N\}$ is used to determine across all trained HMMs which is the most accurate for the tested human action.

HMMs, which have been recently applied with particular success to speech recognition, are a kind of stochastic state transit model [20]. HMMs use observation sequence to determine the hidden states. We suppose $O = \{O_1,O_2,\ldots.O_N\}$ as the observation of the stochastic sequence. HMM with N state is specified by three groups of parameters: $\beta=\{A,B,\pi\}$,where $A=\{a_{ij},=pr(q_t=s_j|q_{t-1}=s_i)\}$ denotes the state transition probability matrix, used to describe the state transition between probability, where, a_{ij} is the

probability of transiting from state q_i to state $q_{j,}$ and $B=\{b_j(k)=pr(v_k|q_t=s_j)\}$, is the matrix of observation probabilities, used to describe the state j, the probability of the output corresponding to the observed values $b_j(k)$ of output symbol v_k at state q_j, and $\pi=\pi\{\pi_i=pr(q_1=s_i)\}$ the initial state probability used to describe the observed sequence of probability when t=1.

Each state of the HMM stochastically outputs a symbol. In state s_i, symbol v_k is output with a probability of $b_i(k)$. If there are M kinds of observation symbols, $b_j(k)$ becomes an N x M matrix, where N is the number of states in the model. The HMM outputs the symbol sequence $O = O_1,O_2,. . . , O_Y$ from time 1 to T. The initial state of the HMM is also stochastically determined by the initial state probability π.

To recognize the observed symbol sequences, we create a single HMM for each action. For a classifier of C actions, we choose the model which best matches the observations from C HMMs $(\beta i=\{Ai, Bj, \pi i\})$, i = 1 . . . C. This means that when a sequence of unknown category is given, we calculate $Pr(\beta_i |O)$ for each HMM βi, and select βc˜. For instance, given the observation sequence $O = O_1,. . . O_T$ and the HMM β_i, according to the Bayes's rule, the problem is how to evaluate $Pr(O|\beta_i)$, the probability that the sequence was generated by HMM β_i, which can be solved using the forward algorithm. Then we classify the action as the one that presents the largest posterior probability

$$c\tilde{}=\text{argmax}_i(Pr(\beta_i|O)). \quad (6)$$

where i indicates the likelihood of test sequence for the ith HMM.

5 Experiments

We evaluate the performance of our algorithm with the public G 3D dataset collected by Bloom et al.[19]. In addition, we evaluated the algorithm with the MSR Action 3D dataset collected by Li et al.[9] and we compared our results with results reported in [9].

Table 1. The subsets of actions used with the MSR Action 3D dataset

Action Set 1 (AS1)	Action Set2 (AS2)	Action set3 (AS3)
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side boxing	Pickup & throw

Table 2. Recognition rates of our method on the G3D action dataset. Results are compared with Bloom et al. [19].

Action Category	Bloom et al.	Our method
Fighting	70.46%	79.84%
Golf	83.37%	100%
Tennis	56.44%	78.66%
FPS	53.57%	54.10%
Driving a car	84.24	81.34%
Misc.	78.21%	89.40%
Overall	71.04%	80.55%

Table 3. Recognition rates of our method on the MSR Action 3D dataset. Results are compared result with Li et al. [9].

Action subset	Li et al.	Our method
AS1	72.9%	86.30%
AS2	71.9%	65.40%
AS3	79.2%	77.70%
Overall	74.7%	76.46%

5.1 Experimental Results

The results of our approach with the G3D dataset collected by Bloom et al.[19], containing 22 types of human actions are summarized in table 2. Each action was performed by 10 individuals for 3 times. Note that we only used the information from the skeleton for action recognition in our algorithm. The experiment was repeated 20 times, and the averaged performance is reported in Table 2. The set of clusters was fixed to K=80, and the number of states to N=6. Half of the subjects were used for training and the rest of the subjects were used for testing. Across experiments, the overall mean accuracy is 80.55% demonstrating that our method performs better recognition than Bloom et al [19].

We also tested our algorithm on the public MSR Action3D database that contains 20 actions. We divided the actions into three subsets (similar to [9]), each comprising 8 actions (see table 1). We used the same parameter settings as previously described. In this test, half of the subjects were used for training and the rest of the subjects were used for testing. Each test was repeated 20 times, and the averaged performance is given in Table3. We compared our performance with Li et al[9]: our algorithm achieves considerably better recognition rates than Li et al.

6 Conclusion

This paper presents a framework to recognize human action from sequences of skeleton data. We use 3D joints positions inferred from skeleton data as input. We propose a method for postures representation that involves joint angles in xy and zy planes within a modified action coordinates system as description of postures. In order to classify action types, we model sequential postures with HMMs. Experimental results illustrate the performance of the proposed method, and also refer to a promising approach to perform recognition tasks using 3D points.

7 References

1. A. Bobick, and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans*, PAMI, 23(3): 257-267), 2001.
2. H. Meng, N. Pears, and C. Bailey. A human action recognition system for embedded computer vision application. In *Proc*. *CVPR*, 2007.
3. J. W. Davis and A. Tyagi. Minimal-latency human action recognition using reliable-inference. *Image and Vision Computing*, 24(5):455-473, 2006.
4. D.–Y. Chen, H.–Y. M. Liao, and S.–W. Shih. Human action recognition using 2-D spatio-temporal templates. In *Proc ICME*, pages 667-670, 2007.
5. V. Kellokumpu, M. Pietikainen, and J. Heikkila. Human activity recognition using sequences of postures. In *Proc IAPR Conf. Machine Vision Applications*, pages 570-573, 2005.
6. C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *Proc ICCV*, Vol. 2, pages 808-815, 2005.
7. J. Zhang and S. Gong. Action categoration with modified hidden conditional random field. *Pattern Recognitoin*, 43: 197-203, 2010.
8. W. Li, Z. Zhang, and Z. Liu. Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1499-1510, 2008.
9. W. Li, Z. Zhang, Z. Liu, "Action recognition based on a bag of 3D points", *CVPRW*, 2010
10. V. M. Zatisiorsky. Kinematics of Human motion. *Human kinetics Publisher*.
11. E. Yu, and J. K. Aggarawal. Human action recognition with extremities as semantic posture representation. In proc. *CVPR*, 2009.
12. L. Xai, C. C. Chen, and J. k. Aggarwal. View invariant human action recognition using histogram of 3D Joints. *2nd international Workshop on Human Action Understanding from 3D Data in conjunction with IEEE CVPR*, pp. 20-27, 2012.
13. M. Z. Uddin, N. D. Thang, J.T. Kim and T.S. Kim, Human Activity Recognition Using Body Joint-Angle Features and Hidden Markov Model. *ETRI Journal*, vol.33, no.4, Aug., pp.569-579, 2011.
14. J. Yamato, J. Ohya, and K. Ishii. Recognition Human action in time-sequential images using hidden markov model. *IEEE int. conf. computer vision pattern recognition*, pp 379-385, 1992.
15. http:// dipresec.king.ac.uk/G3D
16. F. niu and M. abdel-mottaleb. View-invariant human activity recognition based on shape and motion features. *IEEE 6th int. symp. Multimedia software eng*., pp, 546-556, 2004.

17. M. Z. Uddin et al. Human activity recognition using independent components features from depth images. *5th int. conf. ubiquitous healthcare*, pp. 181-183, 2008.
18. M. Z. Uddin, JJ. Lee, and T.-S. Kim, independent shape component-based human activity recognition via hidden markov model*, Appl. Intellig*., vol. 33, no.2, pp. 193-206, 2009.
19. V. Bloom, D. Makris, V. Argyriou, G3D: a Gaming action dataset and real time action recognition evaluation framework, *3rd IEEE, inter., workshop on computer vision for computer games, CVCG*. 2012.
20. L. R. Rabiner . A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. Of the IEEE*, 77(2), pp. 257-285, 1989.
21. N. D. Thang et al., Estimation of 3D Human Body Posture via Co-registration of 3D Human Model and sequential stereo information. *Applied Intell., DOI*: 10.1007/s10489-009-0209-4, 2010.
22. P. Turaga, R. Chellapa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.
23. J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. In *ACM Computing Surveys*, 2011.
24. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, Real-Time Human Pose Recognition in Parts from a Single Depth Image, in *CVPR*, *IEEE*, June 201.
25. T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.
26. D. Weinland, E. Boyer and R. Ronfard. Action recognition from arbitrary views using 3D exemplars*, ICCV* 2007.