Proceedings of the 33rd Chinese Control Conference
July 28-30, 2014, Nanjing, China

Human Activities Segmentation and Location of Key Frames Based on 3D Skeleton

WANG Yongxiong, SHI Yubo

Key Laboratory of Modern Optical System, and Engineering Research Center of Optical Instrument and System, Ministry of Education, University of Shanghai for Science and Technology, Shanghai 200093
E-mail: wyxiong@usst.edu.cn

Abstract: Human activities consist of multiple simple actions, and the temporal information benefit action recognition at all time scales. Considering energy information of human action as action similarity criterion, we present a temporal segmentation method which action videos are firstly segmented to atomic actions based on kinematics information of human skeleton, then the atomic action units are iteratively incorporated in meaningful group by considering similarity of energy information. And the key frames are located at sphere of maximum energy information. We tested our method on two challenging datasets and its performance is better than other state of the art methods.

Key Words: Temporal Segmentation, Human Activity Recognition, 3D Perception, Energy Information

1 Introduction

Human action recognition is one of the popular topics in compute vision as lots of video cameras are widely used. The temporal segmentation is efficiently performed in a sequence of continuous actions are very important for accurately understanding human actions [1-3]. Moreover, many video data may not have exact starting time and ending time of an action and may include the large numbers of null classes. Such action videos are only considered as weakly labeled data [4] for training, and correspondingly, they may lead to extra work for action recognition. Therefore automatic video segmentation or change-point analyses within a sequence of temporal observations become a focused problem of human action recognition.

Microsoft Kinect is an affordable sensor device which can synchronously provide color image and precise depth data (RGB-D). It has been widely used in many applications, such as object tracking [5], indoor 3D mapping [6], event recognition [7] and so on. Human skeleton which represents as body part locations can be rapidly acquired from its depth data [8]. The human actions consist of the sub-actions of human skeleton parts. Thereby it is convenient and efficient for human action recognition using a RGB-D sensor. In this paper, we propose an algorithm of temporal segmentation in order to partition the frames representing the human sub-actions, as well as acquire the key frames in each sub-action groups of frames based on RGB-D date.

Probabilistic models, such as conditional random field [9], hidden markov models [10], is most often used to dealt with human action recognition and temporal segmentation. In some work of these, they partition actions into "hidden states" that correlate to complete motion fragments. The performances based on these methods are promising and competitive. However, most of methods only are applied in small periods of time, where temporal segmentation is not difficult. The methods of bottom-up temporal segmentation based on dynamic Bayes nets are widely used in the speech and natural language processing, but they have much lower accuracy for action recognition [4]. And high-dimensional minimization problem will bring the computation cost is exponential to search the optimal segmentation. The generative approaches are often difficult to handle the unconscious action or the background clutter in long video sequences. Dynamic programming algorithms are employed to solve the inference problem of segmentation and recognition in most methods as well as change-point detection in time series [11]. However, these approaches are difficult to handing null classes or the unconscious actions.

In [12], skeleton based human actions are segmented based on two states of actions: stationary and motion. The action between two stationary states is regarded as a sub-action. Then all human action can be composed of one or more such atomic sub-actions. This method is simple and efficient for simple action, such as walking, waving, and so on. However, because of the ambiguity of human action, only analyzing the human actions is insufficient to understanding complex human actions. An efficient method is modeling both the human activities and the interactions of human-object, human-human and object–object [3, 9]. In particular, the interactions can be easily obtained when the depth information is incorporated in 2D image data. The segmentation of such temporal variations is still a challenging problem in the case. Firstly, the temporal variations may be uncertainty and ambiguity. The next human sub-action might have begun before the previous sub-action ended. Secondly, the complex interactions of human-object and object–object are difficult to approximate with a linear dynamical system and can't use dynamic programming techniques to find the optimal segmentation [3].

In our paper we propose a temporal segmentation approach which considers both the kinematics information of human skeleton and the interactions of human–object in time series. Our approach consists of two steps: the atomic action units are firstly segmented based on kinetic energy of human skeleton, then the atomic action units are iteratively incorporated in meaningful sub-action segments by considering the human-object interactions and gesture of human actions. And we try to locate the temporal key frames

*This work is supported by Innovation Project of Shanghai Graduate Education. No. SHGEUSST1301.

in each meaningful action unit which include more information for action recognition.

The rest of the paper is organized as follows. Related work is reviewed in Section 2. The proposed temporal segmentation method is describes in Section 3. Section 4 reports the experiments on two datasets. Finally we conclude the paper and plan the work for future research in Section 5.

2 Related Work

There is much research in performing human action recognition on the short videos. They suppose the temporal segmentation has been done before action recognition. Complex action recognition in long videos would benefit from accuracy temporal group [3, 9, 13]. [13] discussed the importance of temporal segmentation for overall accuracy of human action recognition and partitioned the video to improve the recognition performance [13]. Clustering based on similarity measure is used to segment video sequences of human action in [2]. They consider activities as long-term temporal objects which are characterized by spatio-temporal features at multiple temporal scales. [14] performed a sequence of change-point analysis in sliding windows running along the time series.

Recently, with the extensive use of inexpensive RGB-D sensors, there are many works in implementing human actions recognition from RBG-D data. In [15] a hierarchical maximum entropy Markov model is used in human activity composed by a set of sub-activities from RGB-D videos. They consider the sub-activities based on human pose and motion as hidden nodes in their model. In [12] they do the temporal segmentation of sub-actions by using new features which include the location and velocity of human skeleton's joints. In [3], before action recognition, they used three methods to partition the temporal series. The three methods include uniform segmentation of fixed size, graph segmentation based on Euclidean distances between the skeleton joints and a segmentation method based on rate of change of the Euclidean distance, which consider the movements of the skeleton joints are smooth or sudden. At last, they use optimization method to combine the three methods. It had high complexity. In [16] they use a discriminative temporal alignment method called maximum margin temporal warping to align two actions sequences. However, they don't consider the context information of human-object interactions which provide useful information for segmenting temporal series of complex activities.

3 Modeling Complex Temporal Composition

From Kinect SDK, we can conveniently obtain the human skeleton which is composed of 19 sticks and 20 joints, such as hand, neck, torso, left shoulder, left elbow, left palm, right shoulder, and so on, as shown in Fig. 1.

Our method has following several steps: We first acquire the locations of the joints of human skeleton using Openni's skeleton tracker from the color and depth images. And we calculate velocity of the joints human skeleton. We perform the atomic temporal segmentation using the kinematics information of human skeleton. There may be much noise in the data because human body parts may be occluded by themselves or some object in real-world case. We secondly search 3D bounding boxes around the human skeleton for detecting the object of interaction with human body. Note that we just need to know the locations of the objects instead of knowing what the objects are. Thirdly, in order to improve the accuracy of temporal segmentation, we use energy function to model the mutual context between the location of object and joints of the human skeleton. Finally, we incorporate the atomic action into complete meaningful action segments by similarity of energy.

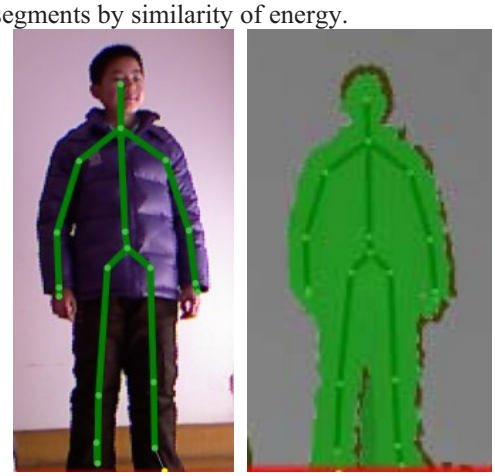

Fig. 1: The human skeleton consisted of sticks and joints

3.1 Minimal Temporal Segmentation Based on Kinetic Energy of Human Skeleton

Human action has the two states, i.e. stationary and motion states. We can consider that it is a sub-action between two stationary states. The human activity consists of the atomic sub-action. The velocity and kinetic energy changes of human actions have similar periodic regularity. Thus we simply use kinetic energy of the joints of human skeleton to denote motion state of human skeleton.

To obtain the kinetic energy of the joints of human skeleton, first, we acquire the x, y, and z coordinates of the joints. Then calculate the velocity and kinetic energy of all joint of human skeleton as follows,

$$\begin{aligned} E_{k,t} &= \sum_{i=1}^{n} k_i v_{i,t}^2 = \frac{1}{\Delta t^2} \sum_{i=1}^{n} \boldsymbol{K}^T \left(\left| \boldsymbol{P}_{i,t} - \boldsymbol{P}_{i,t-1} \right| \right)^2 \\ &= \frac{k_i}{\Delta t^2} \sum_{\phi(x_{i,t}, y_{i,t}, z_{i,t}) \in F_t} (x_{i,t} - x_{i,t-1})^2 + (y_{i,t} - y_{i,t-1})^2 + (z_{i,t} - z_{i,t-1})^2, \end{aligned} \tag{1}$$

where $E_{k,t}$ is the overall kinetic energy of the joints of human body at frame F_t, and $v_{i,t}$ is the velocity of ith joint at frame F_t. Φ is the set of all joints of human skeleton, the subscript i represents ith joint (i=1,…, 20). $P_{i,t}(x_{i,t}, y_{i,t}, z_{i,t})$ is the position of ith joint at the frame F_t. $\boldsymbol{K}$ is a parameter column vector. Generically, the overall velocities of segmentation points are not always 0 in equality (1). However, the velocity at start or end moment is always lowest. And the kinetic energy reaches the peak within each group of video.

The smoothed curve of kinetic energy of a long action, detected segmentation points and key frames are shown in Fig.2. There may are serious shake and too many local minimum in the curve of kinetic energy. This may be caused by noise or unconscious motion. Thus it is need to smooth the curve of kinetic energy. Although the method can avoid merging two sub-activities into one segment, these would result in over-segmentation for the complex sub-action

recognition. For example, we may briefly stop in the course of picking up a book. Therefore, we consider modeling similarity of adjacent segments for merging two similar segments. Moreover, modeling interactions with objects can provide useful information for recognizing complex activities.

3.2 Determining Key Frame in a Temporal Segment

After performing the minimal temporal segmentation using the kinematics information of human skeleton, video sequence is partitioned into many short temporal segments of variable length. [9] used the motion segment classifiers to locate each segment by measuring image-based similarities on the basic of such temporal decomposition results. In contrast, we use energy function to represent the similarity information of human action. We consider that the actions contain the information of potential energy besides the kinetic energy. Thus, to represent the action similarities we model the energy functions of human action which include both potential energy of joints of human skeleton and potential energy of objects interacted with human body.

To acquire the coordinates of objects, we only consider the objects that locate in 3D bounding boxes around the skeleton. This improves the accuracy of detections as well as real time of detection. We then carry out a simply threshold based object detectors on the cloud RGB image. We can obtain the exact x, y and z coordinates of the detected objects.

After acquiring the locations of the objects inside the 3D bounding box of human skeleton, we model potential energy of human–object interactions and gesture of human body for temporal segmentation. In view of the consistency of action in the adjacent temporal segmentations, we build two energy functions to respectively represent potential energy of human body and potential energy of objects which interact with human body. They are defined as follows:

$$\begin{aligned} E_{AP,t} &= \boldsymbol{L}^T(\boldsymbol{P}_{i,t} - \boldsymbol{P}_{i,0}) \\ &= \sum_{\phi \in F_t} l_i[(x_{i,t} - x_{i,0}) + (y_{i,t} - y_{i,0}) + (z_{i,t} - z_{i,0})], \end{aligned} \tag{2}$$

$$\begin{aligned} E_{OP,t} &= \sum_j \boldsymbol{M}^T(\boldsymbol{P}_{i,t} - \boldsymbol{P}^O_{j,t}) \\ &= \sum_j \sum_{\phi \in F_t} m_i[(x_{i,t} - x^O_{j,t}) + (y_{i,t} - y^O_{j,t}) + (z_{i,t} - z^O_{j,t})], \end{aligned} \tag{3}$$

where $E_{AP,t}$, $E_{OP,t}$ are potential energy of human body and potential energy of objects respectively. $P_{i,0}(x_{i,0}, y_{i,0}, z_{i,0})$ is the initial position of ith joint of human skeleton. It can be considered as the natural gesture of human, such as the natural standing gesture of human shown in Fig. 1. If the initial positions of joints are difficult to be acquired, we can use the mean value of position of ith joint. $P^O_{j,t}(x^O_{j,t}, y^O_{j,t}, z^O_{j,t})$ is the position of jth object which interacts with human body (j=1,…,). $\boldsymbol{L}$ and $\boldsymbol{M}$ are the parameter column vectors respectively. $E_{AP,t}$ represents the deviation of gesture position between the current position of joints and initial position of joints (or mean value of positions of joints). $E_{OP,t}$ represents the context from object interactions along with human pose.

Based on above the definition, we consider the key frames of human actions should contain more information, i.e. maximum energy. To confirm the key frames centers for action recognition, we first find out the frames (blue point as shown in Fig. 2) which have local maximum kinetic energy at a time interval. Then we calculate the sum of kinetic and potential energy, and consider the frame with maximum total energy is the key frame at a time interval.

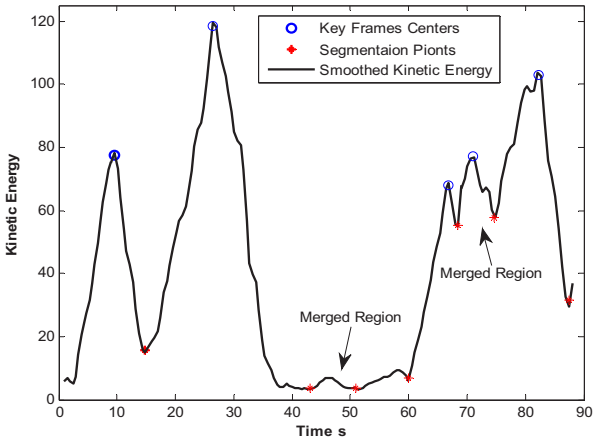

Fig. 2: The smoothed curve of kinetic energy of a long action and detected segmentation points and key frames (k_i=10).

3.3 Segments Merging by Similarity Function of Energy

To avoid over-segmentation of the complex sub-action recognition, we consider merging atomic adjacent segments into meaningful groups by similarity function of energy as shown in Fig. 2. We define the similarity of energy between frame T_1 and frame T_2 as follows,

$$Sim(t_1, t_2) = \frac{(E_{AP,t1} + E_{OP,t1}) - (E_{AP,t2} + E_{OP,t2})}{E_{AP,t1} + E_{OP,t1}}, \tag{4}$$

The merging algorithm starts with the minimal temporal segmentation obtained in Section 3.1. Each iteration combines two adjacent segments when the similarity function of two key frames in two adjacent segments is less than the threshold, or else keeps two segments. The merging algorithm implicitly assumes that the similarities of all frames in each segment are less than the threshold. Thus the threshold is an important parameter for grouping temporal segmentation.

4 Experiments

To evaluate our temporal decomposition approach, we do some experiments on two datasets. First, we test the performance of our model to simple actions on Microsoft Research Cambridge-12 (MSRC-12) gesture dataset [17]. Second, the effectiveness of our segmentation method to complex actions is verified on Cornell Activity Dataset-1200 (CAD-120) [3].

Fig. 3: The RGB-images of activity of *making cereal* which include four sub-activities: moving, reaching, pouring and placing.

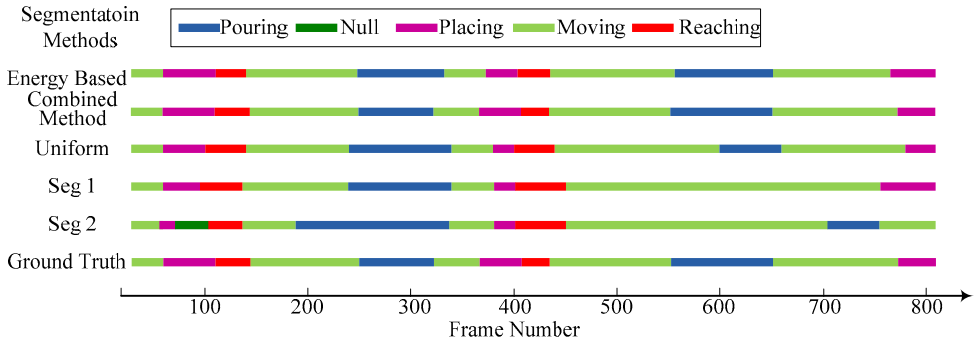

Fig. 4: The segmentation comparison of the sub-activity of various methods. Note that the results come from [3] expect our energy based method.

4.1 Temporal Segmentation of Simple Action

The MSRC-12 gesture dataset comprises of sequences of human skeletal body part movements using the Kinect Pose Estimation pipeline. It consists of 594 sequences, 719359 frames collected from 30 people performing 12 gestures. In total, there are 6244 gesture instances. The ground-truth temporal segmentation is given by manual.

We use the representation based on kinetic and potential energy in the experiments. In order to get the segmentation points and key frames, we search the local maximal and minimal points of kinetic energy at given time intervals (2 second to 6 second). The local minimal points of kinetic energy are the segmentation points of atomic segments and the domain with local maximal energy are used to compare the similarity for merging the adjacent segments. In the test, we restrict the maximal duration of a group is less than 20 seconds. $\boldsymbol{K}=[10,\ldots,10]^{T}$, and all elements of parameter vectors $\boldsymbol{L}$, and $\boldsymbol{M}$ are set to 1. The threshold of similarity is set to 0.85. The domains which can be merged are pointed out in Fig.2. The average precisions of temporal segmentation are shown in Table 1. The experiments show that our method is promising.

Table 1: The average precisions of temporal segmentation

Description	Precisions of segmentations
Crouch or hide	93.3%
Put on night vision goggles	95%
Shoot a pistol	92.6%
Throw an object	96%
Change weapon	94.6%
Kick	96.1%

4.2 Temporal Segmentation of Complex Action

To test the ability of our model to group the temporal structure of complex activities, in the experience we use a high-level activity dataset called Cornell Activity Dataset-1200 (CAD-120) [3], which contains 120 activity sequences of ten different high-level activities performed by four different subjects. There are a total of 61,585 RGB-D video frames in this dataset. The high-level activities includes making cereal, stacking objects, picking objects, cleaning objects, taking food, and so on. It is suitable for testing the complex temporal segmentation method.

To compare to the result of [3], we segment the high-level activity video of *making cereal* as shown in Fig.3. The *making cereal* activity involves four sub-activities: moving, reaching, pouring and placing, which are indicated in Fig. 3. Note that we just employ temporal segmentation of the videos instead of recognizing the sub-activities. The comparison of subactivity segmentation by various methods is shown in Fig. 4 where our results of segmentation, the results based on various other segmentations methods, and the ground-truth segments of *making cereal* activity are shown. Other methods include uniform segmentation, image based segmentation, a method based on the rate of change of the Euclidean distance, and combined based method [3]. It can be seen that almost all segmentation methods have some errors from Fig. 4.

Figure 4 shows our proposed algorithm for combining the segments based on information of energy can effectively group the meaningful segments for more complex activities that are considered as a composition of shorter or simpler actions, for example cooking. And we acquire the best precisions for segment the temporal sequence of simple but periodic actions, such as walking, waving, and boxing. Our segmentation method based on energy modelling is competitive although its segmentation accuracy is slightly lower than the best result reported in [3], which is combined in three methods by complex optimization.

5 Conclusion and Future Work

In the paper we present a method to segment the temporal structure based on kinetic and potential energy of human 3D skeleton. First we find out the atomic temporal segment point based on local minimal kinetic energy. Second we determine key frame according to maximal total energy in each segment. Finally, we merge the atomic temporal segments to meaningful group based on energy similarity criterion. This is beneficial to recognize both complex human activities and simple actions. Future work will continue incorporating other methods and representations of actions to recognize complex human activities semantically.

References

[1] K. Li, J. Hu, and Y. Fu, Modeling complex temporal composition of actionlets for activity prediction, In *Proceedings of 12th European Conference on Computer Vision (ECCV)*, 2012: 286–299.

[2] M. Hoai and F. De la Torre, Maximum margin temporal clustering. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2012: 1-9.

[3] H. Koppula, R. Gupta, and A. Saxena, Learning human activities and object affordances from RGB-D videos, *The International Journal of Robotics Research*, 32(8): 951–970, 2013.

[4] M. Hoai, Z.-Z. Lan, and F. De la Torre, Joint segmentation and classification of human actions in video. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2011: 3265-3272.

[5] M, Krainin, P. Henry, X. Ren and D. Fox, Manipulator and object tracking for in-hand 3D object modeling, *The International Journal of Robotics Research*, 30: 1311–1327, 2011.

[6] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, RGB-D Mapping: Using depth cameras for dense 3D modeling of indoor environments. In *Proceedings of 12th International Symposium on Experimental Robotics (ISER)*, 2010.

[7] B. Fosty, C. F. Crispim-Junior, et al, Event Recognition System for Older People Monitoring Using an RGB-D Camera, In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.

[8] J. Shotton, A. Fitzgibbon, M. Cook, et al, Real-time human pose recognition in parts from single depth images, In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2011: 1297–1303.

[9] J. C. Niebles, C.-w. Chen, and L. Fei-fei, Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In *Proceedings of 12th European Conference on Computer Vision (ECCV)*, 2012.

[10] K. Li, J. Hu, and Y. Fu, Modeling Complex Temporal Composition of Actionlets for Activity Prediction. In *Proceedings of 12th European Conference on Computer Vision (ECCV)*, 2012, 286–299.

[11] X. Xuan and K. Murphy, Modeling changing dependency structure in multivariate time series. In *Proceedings of 24th International Conference on Machine Learning,* 2007.

[12] H. Shuzi, Y, Jing, and C. Huan, Human actions segmentation and matching based on 3D skeleton model, In *Proceedings of 24th China Control Conference*, 2013: 5877-5882.

[13] S. Satkin and M. Hebert, Modeling the temporal extent of actions. In *Proceedings of 11th European Conference on Computer Vision (ECCV)*, 2010.

[14] Z. Harchaoui, F. Bach, and E. Moulines. Kernel change-point analysis. In *Neural Information Processing Systems*, 2009.

[15] J. Sung, C. Ponce, B. Selman and A. Saxena, Unstructured human activity detection from RGBD images, In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2012: 842-849.

[16] J. Wang and Y. Wu, Learning maximum margin temporal warping for action recognition，In *Proceedings of European Conference on Computer Vision*, 2013: 2688-2695.

[17] S. Fothergill, H. M. Mentis, S. Nowozin, and P. Kohli, Instructing people for training gestural interactive systems, In *Proceedings of ACM Conference on Computer-Human Interaction*, 2012: 1737--1746.