A Novel Unsupervised Method for Temporal Segmentation of Videos

Xiangbin Shi[1,2,*], Yaguang Lu[1], Cuiwei Liu[2], Deyuan Zhang[2] and Fang Liu[2]
[1] *School of Information, Liaoning University*
Shenyang, P. R. China
[2] *School of Computer Science, Shenyang Aerospace University*
Shenyang, P. R. China
[*] *Email: sxb@sau.edu.cn*

***Abstract*—In this paper, we aim to address the problem of temporal segmentation of videos. Videos acquired from real world usually contain several continuous actions. Some literatures divide these real-world videos into many video clips with fixed length, since the features obtained from a single frame cannot fully describe human motion in a period. But a fixed-length video clip may contain frames from several adjacent actions, which would significantly affect the performance of action segmentation and recognition. Here we propose a novel unsupervised method based on the directions of velocity to divide an input video into a series of clips with unfixed length. Experiments conducted on the IXMAS dataset verify the effectiveness of our method.**

***Keywords*-temporal video segmentation, action segmentation, action recognition**

I. Introduction

The fast development of the video capture technology has created a great need for methods of intelligent video analysis. Analyzing and understanding human actions in videos is one of the hot topics.

Many traditional action segmentation and recognition methods aim to recognize the actions in manually segmented videos. These methods are executed frame by frame on the pre-segmented videos which contain only one action. But the videos we get from the real world usually contain several continuous actions and the videos can be quite long. Some methods, like [1], [2], divide input videos into sequences of fixed-length video clips before executing action segmentation and recognition since the actions can be better described with the features obtained from video clips. However, a fixed-length video clip might contain frames of different actions, since the boundaries of actions cannot always just hit the boundaries of video clips. That will significantly affect the performance of action segmentation and recognition.

In order to solve this problem, we propose a novel unsupervised temporal video segmentation method called Main Direction Segmentation Method to split a long-term video into a series of video clips. What we want to get are video clips containing frames of only one action. As Rubin & Richards [3] have proposed, it is an effective way to distinguish different motions by detecting moving boundaries. And Briassouli et al.[4] found that motion boundaries in video are usually appearing with sequential changes. Both [3] and [4] are suitable for videos with similar motion intensity actions, since they are quite sensitive to the intensity of the actions. Our method is also a boundary-based method. This method performs video segmentation according to the direction of movement. But unlike the method in [4], our method can be executed on videos with multiscale actions since the velocity directions are not directly related to the motion intensity.

II. Our Approach

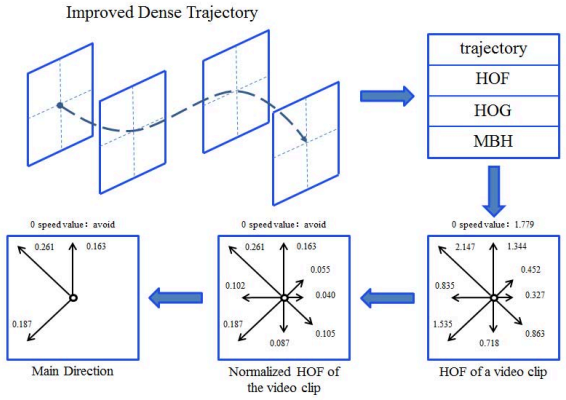


Figure 1. IDT and main direction. The top half of this figure is the instruction of IDT, and the bottom half is the major idea of main direction.

In this paper, we propose a novel unsupervised segmentation method called Main Direction Segmentation Method. It is a boundary-based automatic temporal segmentation method, and it achieves segmentation with the main direction of movements. The main directions are achieved by computing the velocities of movements which are described with Improved Dense Trajectory(IDT)[5]. The velocity of movements in a frame can be projected to eight sub-velocities with different directions. Input videos are divided into clips by the distance of the main velocity vectors of adjacent frames, and we hope that all frames in a clip belong to the same action.

978-1-5386-2636-8/19/$31.00 ©2019 IEEE
DOI 10.1109/ICVRV.2017.00017

The features of video X can be extracted as a series: $T = \{t_1, t_2, \cdots, t_n\}$. Here t_i is a trajectory which including four different types of descriptors: the trajectory feature, Histogram of Oriented Gradients(HOG), Histogram of Optical Flow(HOF) and Motion Boundary Histogram(MBH). Supposing $T_k = \{t_{k1}, t_{k2}, \cdots, t_{km}\}$ are the trajectories through frame k, we can get velocity information of k with its HOF. HOF projects the velocity of a trajectory to eight sub-velocities in different directions. With these sub-velocities, the velocity vector of a trajectory can be shown as $\{v_{k0}, v_{k1}, v_{k2}, \cdots, v_{k8}\}$. Then, we describe the directions of human movements in that frame with several maximum sub-velocities. The amount of these maximum sub-velocities is k_x and the series of these sub-velocities is called as main direction.

III. Experiments

In this section we evaluate the effectiveness and practicability of the Main Direction Segmentation Method on IXMAS human action dataset. This dataset has 5 cameras, and we only use camera 0 to exam our method. The leave-one-out(LOO) test is adopted here. We extract the IDT features of the videos , and then divide these videos into video clips by Main Directions.

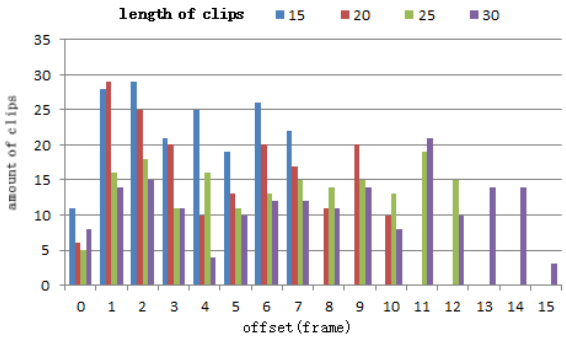
The offsets show the amount of frames between the divided action boundaries and the real action boundaries. The offsets of action boundaries between Main Direction Segmentation Method and the method with fixed-length clips in [6] are compared in frame level. A series of video clips in different lengths are built, while the maximum length of our method is set correspondingly. There are 181 action boundaries on camera 0, and all these boundaries are used to evaluate the distribution of their offsets in each length and their average offsets. The results are shown in Figure 2. It can be seen that the offsets of the method in [6] are almost uniformly distributed, while our method is able to obtain much fewer offsets. Comprehensively, our Main Direction Segmentation Method can effectively reduce the amount of the error segmentation compared to methods based on fixed-length clips.

IV. Conclusion

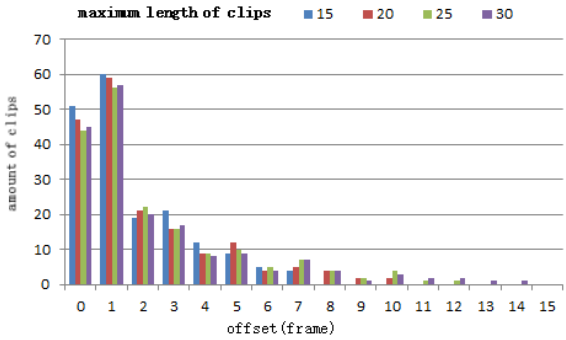
We have presented an unsupervised temporal segmentation method in this paper. The proposed Main Direction Segmentation Method can divide a long-term video into a sequence of clips reasonably. Experiments on IXMAS human action dataset proved the effectiveness of our method.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China(NSFC) under Grant No.61602320 and No.61170185, Liaoning Doctoral Startup Project under Grant No.201601172, Foundation of Liaoning Educational Committee under Grant No.L201607, and the Young Scholars Research Fund of SAU under Grant No.15YB37.



(a) Offsets of the method in [6]



(b) Offsets of our method

Figure 2. The distributions of the offsets of the method in [6] and our method are shown in subfigure (a) and (b).

References

[1] Y Cheng, Q Fan, S Pankanti, and A Choudhary. Temporal sequence modeling for video event detection. In *27th IEEE Conference on Computer Vision and Pattern Recognition: 24-27 June 2014; Columbus, Ohio, USA*, pages 2235–2242. IEEE, 2014.

[2] J Wang, X Nie, Y Xia, Y Wu, and S C Zhu. Cross-view action modeling, learning and recognition. In *27th IEEE Conference on Computer Vision and Pattern Recognition: 24-27 June 2014; Columbus, Ohio, USA*, pages 2649–2656. IEEE, 2014.

[3] Boundaries of visual motion, 1985.

[4] A Briassouli, V Tsiminaki, and I Kompatsiaris. Human motion analysis via statistical motion processing and sequential change detection. *Journal on Image & Video Processing*, 2009, 1:1–16, 2009.

[5] H Wang and C Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision: 3-6 December 2013; Sydney, Australia*, pages 3551–3558. IEEE, 2013.

[6] M Hoai, Z Z Lan, and F D L Torre. Joint segmentation and classification of human actions in video. In *24th IEEE Conference on Computer Vision and Pattern Recognition: 20-25 June 2011; Colorado Springs, Colorado, USA*, pages 3265–3272. IEEE, 2011.