Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 1, January 2018, pp. 101-110
DOI: 10.23940/ijpe.18.01.p11.101110

A Novel Double-Layer Framework for Joint Segmentation and Recognition of Multiple Actions

Cuiwei Liu[a], Yaguang Lu[b], Xiangbin Shi[a,b,*], Deyuan Zhang[a] and Fang Liu[a]

[a]*Computer Science, Shenyang Aerospace University, Shenyang, 110136, China*
[b]*School of Information, Liaoning University, Shenyang, 110036, China*

Abstract

This paper aims to address the problem of joint segmentation and recognition of multiple actions in a long-term video. Since features obtained from a single frame cannot describe human motion in a period, some literatures initially divide a long-term video into many video clips with fixed length and represent a long-term video as a sequence of video clips. However, a fixed-length video clip may contain frames from two adjacent actions, which would significantly affect the performance of action segmentation and recognition. In this paper, we develop a double-layer framework for segmenting and recognizing multiple actions in a long-term video. In the first layer, a novel unsupervised method based on the directions of velocity is proposed to initially divide an input video into a series of clips with unfixed length. The second layer takes a sequence of video clips as input, and employs a joint segmentation and recognition method to group video clips into several segments while simultaneously labeling the action category for each segment. Experiments conducted on the IXMAS action dataset verify the effectiveness of the proposed approach.

Keywords: action segmentation; action recognition; hierarchical framework

(Submitted on October 2, 2017; Revised on November 15, 2017; Accepted on December 10, 2017)

© 2018 Totem Publisher, Inc. All rights reserved.

1. Introduction

The fast development of video capture technology has created a great need for methods of intelligent video analysis. Analyzing and understanding human actions in videos is a hot and active topic, especially for actions in long-term videos. The two main problems are “What the action is” and “When it happens”.

Traditional action recognition methods divide an input video into coherent constituent action segments at first, and then classify these segments into different action categories. But the loss of action information might happen since the recognition and segmentation of human actions are done separately. Some recent literatures avoid the problem of information loss by executing video segmentation and action recognition jointly, and they are performed at two different granularities. One group of methods [4,11,26,27] is executed in frame level, while another strategy [6,10,28,32] is to initially divide an input video into a sequence of video clips with fixed length before performing action segmentation and recognition. Compared with the frame-based methods, the clip-based methods can describe human actions better, since more comprehensive features can be obtained from a video clip than a single frame. Nevertheless, clips segmented through the equidistant division might contain frames of several different actions, which may affect the performance of human action recognition in videos.

In this paper, we propose a novel double-layer framework to segment and recognize multiple actions in long-term videos by using unfixed-length video clips. An unsupervised temporal segmentation method is proposed to split a long-term video into a series of video clips in the first layer of our framework. This pre-segmentation method performs initial video segmentation according to the direction of movement, and is called Main Direction Initial Segmentation Method. The second layer of our framework takes the series of video clips as input, and employs a joint action segmentation and recognition method to divide a long-term video into several segments while simultaneously labeling the action category of each segment.

* Corresponding author.
E-mail address: sxb@sau.edu.cn

It is worth noting that the action segmentation and recognition method in this layer should be able to work on both frames and video clips, since some video clips may contain only one frame. We find that the method proposed in [11] is an ideal choice. The structure of our double-layer framework is shown in Figure 1. With this framework, we can reduce the disturbance caused by false initial segmentation and get the advantage of video clips.

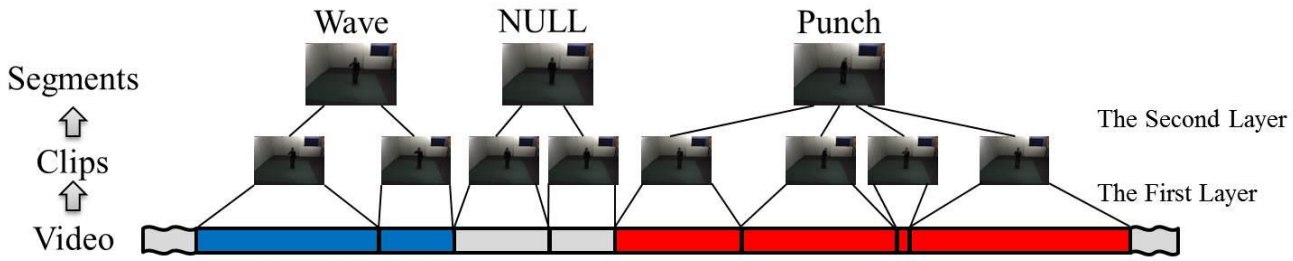

Figure 1. The structure of the double-layer framework.

This paper is organized as follows: Section 2 reviews the related work. Section 3 shows our main contributions, including the Main Direction Initial Segmentation Method and the double-layer framework. Experimental results on the IXMAS action dataset are shown in Section 4. Section 5 is the conclusion of our work.

2. Related Work

Most of the traditional action recognition methods aim to annotate the action categories of pre-segmented video segments, each of which only contains one action. Aggarwal & Cai [1] presented a method of recognizing actions from multi-view cameras with pre-segmented video segments. Polana & Nelson [21] described an action with optical flow fields, assuming that human motion is periodic. Each of these videos contains only one action so that the entire video can be divided into some circular processes of a whole action. Besides, many datasets were established in a similar way, such as KTH dataset [23], Weizmann dataset [2], Hollywood2 dataset [18] and Olympic Sports dataset [19]. Videos in these datasets are manually segmented and only contain one action.

In order to alleviate the laborious and time-consuming manual annotations of action segmentations, some literatures propose to segment multiple actions automatically. Chen et al. [5] proposed a sliding-window based segmentation method by using Fisher Vector Coding. Some other methods process video segmentation with series models. These methods usually describe an input video with a series of features before evaluating the matching degree between the action model and the feature series. Lv & Nevatia [16] achieved automatical segmentation and recognition of human actions by training a set of Hidden Markov Models as weak classifiers of AdaBoost. Lu et al. [15] represented a video with a series of key frame. Syeda-Mahmood et al. [25] proposed a method based on speed value. Li et al. [14] used temporal subspace clustering to depart different actions. Series model is useful but complicated, and using action boundaries to acquire action segments is a much easier way. Marr & Vaina [17] have discussed 3D human movement segmentation problem as early as the 1980s and suggested to separate different movements with resting state. Rubin & Richards [12] and Rui & Anandan [22] proposed that it was an effective way to distinguish different motions by detecting moving boundaries. Ogale et al. [20] checked the maximum and minimum of optical flow within human body contour. Briassouli et al. [3] found motion boundaries in videos with sequential changes. Both [20] and [3] are only suitable for videos with similar motion intensity actions, since they are quite sensitive to the intensity of the actions. Our Main Direction Initial Segmentation Method is also a boundary-based method, and it performs segmentation according to the main direction of movements. Unlike methods in [3,20], our method can be executed on videos with multiscale actions since the velocity directions are not directly related to the motion intensity.

Video segmentation and action recognition are performed separately in many methods, such as [24]. However, these two procedures are closely related. Action recognition is based on the results of video segmentation, while the recognition results also provide clues for video segmentation. So, it is more reasonable to combine these two procedures into a unified framework. Hoai et al. [11] proposed a method to segment and classify actions at the same time using a framework based on multi-class SVM [7], and they segmented a video in a bottom-up structure. The method in [11] is executed in frame level, while the features obtained from a single frame cannot describe human motion well. Lei et al. [13] proposed a hierarchical joint segmentation and recognition framework that extracts features using CNN and describes the sub-action sequences with HMM. Cheng et al. [6] divided a video into fixed-length clips initially, and then performed segmentation and recognition jointly. However, there is a problem that a fixed-length clip might contain frames of several different actions, which will significantly affect the performance of action segmentation and recognition. This problem is caused by the contradiction between the fixed

length of clips and the variable length of actions. In order to solve this problem, we propose a double-layer framework to minimize the contradiction by dividing input videos into unfixed-length video clips before performing joint segmentation and recognition of multiple actions.

3. Our Approach

In this section, our double-layer framework is described in detail. We propose an unsupervised initial segmentation method called Main Direction Segmentation Method, which is executed in the first layer. In the second layer, a joint segmentation and recognition method is employed to achieve segmentation and recognition of multiple actions in a long-term video.

3.1. Double-layer Framework

In our double-layer framework, input videos are firstly divided into a series of video clips. These clips can be obtained by performing an initial segmentation method in the first layer. Then, the second layer employs a joint segmentation and recognition method to classify these clips to corresponding categories of human actions. The structure of the double-layer framework is shown in Figure 2.

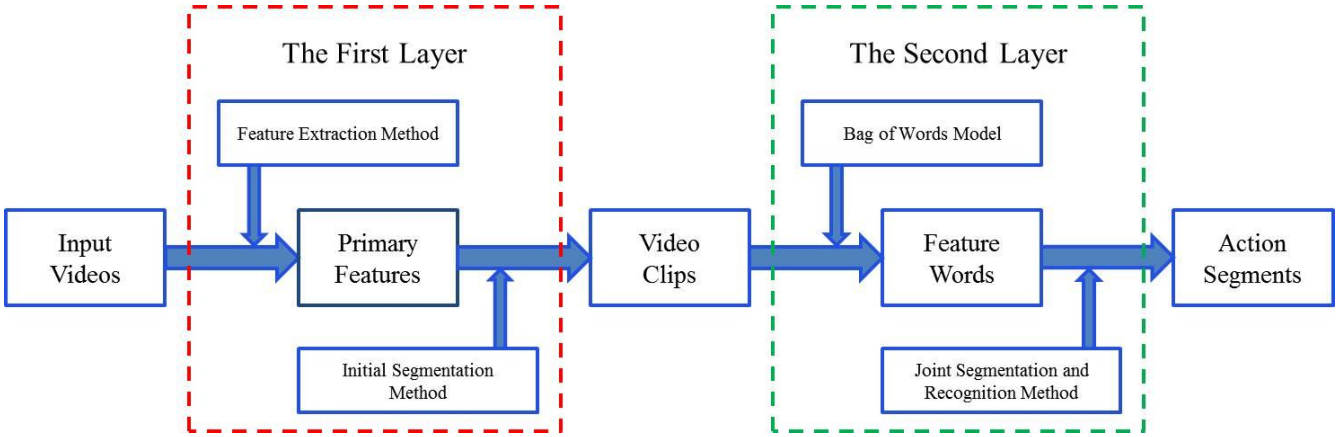

Figure 2. Components of our double-layer framework. In this framework, human action segmentation and recognition are divided into two main stages, and each stage is called a layer.

What we hope to obtain from the first layer are video clips containing frames of only one action, and the second layer takes these video clips as input. Each video clip is described by Bag of words model, and the method presented in [11] is adopted in the second layer for joint segmentation and recognition of multiple actions. Along with the execution of this method, video clips are aggregated to several segments and these segments are classified to corresponding action classes. Our framework can retain the advantages of clip based joint segmentation and recognition method while avoiding the negative effect. As is shown in Figure 2, the double-layer framework is extensible. The modularity structure of the framework makes it possible to be modified and improved.

3.2. Main Direction Initial Segmentation Method

We propose an unsupervised initial segmentation method called Main Direction Initial Segmentation Method. Main Direction Initial Segmentation Method is a boundary-based automatic initial segmentation method and achieves segmentation with the main direction of movements. Main direction is obtained by computing the velocities of movements, which are described with Improved Dense Trajectory (IDT) [31]. The velocity of movements in a frame can be projected to eight sub-velocities indicating different directions. The summation of each sub-velocity is obtained frame by frame, and then the main velocity vector of every frame is built up with the directions of several maximum velocity components. An input video is divided into multiple clips according to the distance of main velocity vectors of adjacent frames. The length of each clip depends on the properties of the actions, and we hope that all frames in a clip belong to the same action.

Given an input video X, a group of IDT features are extracted from X: $T=\{t_1,t_2,\ldots,t_n\}$. Here t_i is a trajectory across 15 frames, which includs four different types of descriptors: the trajectory feature, Histogram of Oriented Gradients (HOG) [8], Histogram of Optical Flow (HOF), and Motion Boundary Histogram (MBH) [9]. Suppose that $T_k=\{t_{k1},t_{k2},\ldots,t_{km}\}$ are trajectories extracted through frame k, the HOG, HOF and MBH of frame k can be obtained by summing the features of T_k. We can get velocity information of k with its HOF. Generally, HOF projects the velocity of a trajectory to eight sub-velocities

in different directions. The angle of two neighboring sub-velocities is 45° , which is the same as optical flow bins. With these sub-velocities, the velocity vector of a trajectory can be shown as $[v_1,v_2,\ldots,v_8]$. Besides, HOF in IDT adds v_0 to show the amount of HOF cells with velocities below the velocity threshold set up by IDT. So, the velocity vector of frame k can be represented as $[v_{k0},v_{k1},v_{k2},\ldots,v_{k8}]$ by summing up the corresponding sub-velocities of T_k. Here, v_{k0} is ignored since the relatively static proportion of the action is not a major concern. Then, we describe the directions of human movements in that frame with several maximum sub-velocities in $\{v_{k1},v_{k2},\ldots,v_{k8}\}$. The amount of these maximum sub-velocities is k_x and the series of these sub-velocities is called as the main direction. To reduce the influence of motion intensity, the main direction should be normalized as is shown in Figure 3.

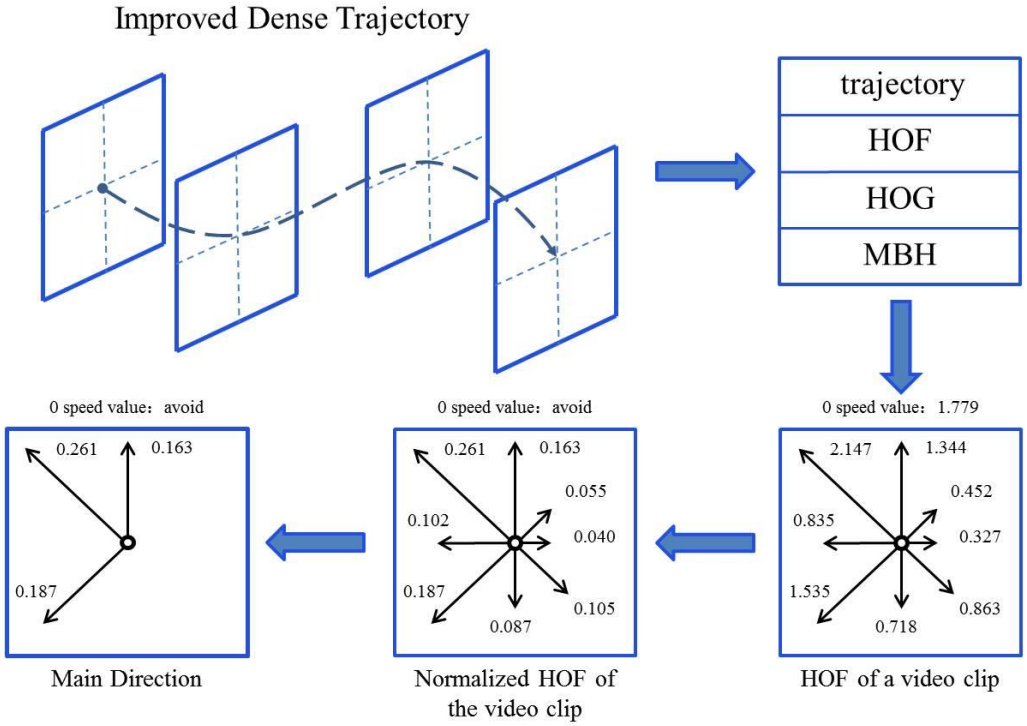

Figure 3. IDT and main direction. The top half of this figure is the instruction of IDT, and the bottom half is the major idea of main direction.

Performing actions separation by checking the changes of the main directions between adjacent frames is effective, since we find that the main directions always change in the boundaries of different actions. $\|v_{k,k+1}\|$ is the segmentation criterion of our method, which is the Euclidian distance between frame k and k+1 as Eq.1 shows. A threshold δ_v is set to adjacent frames, and the maximum length of a video clip is limited to $l_{\max}$. Frame k and frame k+1 are divided into two different video clips when $\|v_{k,k+1}\| \geqslant \delta_v$ or the length of the clip reaches $l_{\max}$.

$$\|v_{k,k+1}\| = \sqrt{\sum_{i}^{8}\left(v_{ki} - v_{(k+1)i}\right)^2} \tag{1}$$

Our method initially divides an input video into a sequence of video clips, which are then described with Bag of words features.

3.3. Joint Segmentation and Recognition

Hoai et al. [11] trained their recognition model by using multi-class SVM [7] and performed segmentation with dynamic programming. With the training time series $\{X^1,X^2,\ldots,X^n\}$ and change points between actions $0=s_1^i<\ldots<s_{k+1}^i=len(X^i)$, they train a model for temporal actions using multi-class SVM and get the weight vector $W=\{w_j\}$.

$$\begin{aligned} &\underset{W_j,\xi_t^i\geq 0}{\text{minimize}}\frac{1}{2m}\sum_{j=1}^{m}W_j^{\,2}+C\sum_{i=1}^{n}\sum_{t=1}^{k_i}\xi_t^i \\ &s.t.\ \left(W_{y_t^i}-W_y\right)^T\varphi\left(X^i_{\left(s_t^i,s_{t+1}^i\right]}\right)\geq 1-\xi_t^i,\ \forall i,t,y\neq y_t^i \end{aligned} \tag{2}$$

The parameter C controls the trade-off between a large margin and less constrained violation. The learned weight vector $W=\{w_j\}$ is used to segment the test time series with SVM scores generated by minimizing Eq.3.

$$\begin{aligned} &\underset{k,s_t,y_t,\xi_t\geq 0}{\text{minimize}}\sum_{t=1}^{k}\xi_t \\ &s.t.\ l_{min}\leq s_{t+1}-s_t\leq l_{max},\ \forall t,s_1=0,s_{k+1}=len\left(X\right), \\ &\left(W_{y_t}-W_y\right)^T\varphi\left(X^i_{\left(s_t^i,s_{t+1}^i\right]}\right)\geq 1-\xi_t^i,\ \forall i,t,y\neq y_t^i \end{aligned} \tag{3}$$

Then the joint segmentation and recognition of actions is performed on the test series $W=\{w_j\}$ with dynamic programming similar to [11]. Using dynamic programing for time series X, we should consider the best segmentation for the truncated time series $X_{(0,u)}$ with Eq.4. The function $\varphi()$ in Eq.4 is the feature mapping in segment level and the length of the series $u\in(0,len(X)]$.

$$\begin{aligned} &f(u)=\min_{k,s_t,y_t,\xi_t\geq 0}\sum_{t=1}^{k}\xi_t \\ &s.t.\ l_{min}\leq s_{t+1}-s_t\leq l_{max},s_1=0,s_{k+1}=u,\ \forall t, \\ &(W_{y_t}-W_y)^T\varphi(X_{(s_t,s_{t+1}]})\geq 1-\xi_t\ \forall t,y\neq y_t \end{aligned} \tag{4}$$

4. Experiments

In order to evaluate the effectiveness of the proposed double-layer framework and the Main Direction Initial Segmentation Method, we conduct experiments on the INRIA Xmas Motion Acquisition Sequences (IXMAS) action dataset [30].

4.1. Experimental setting

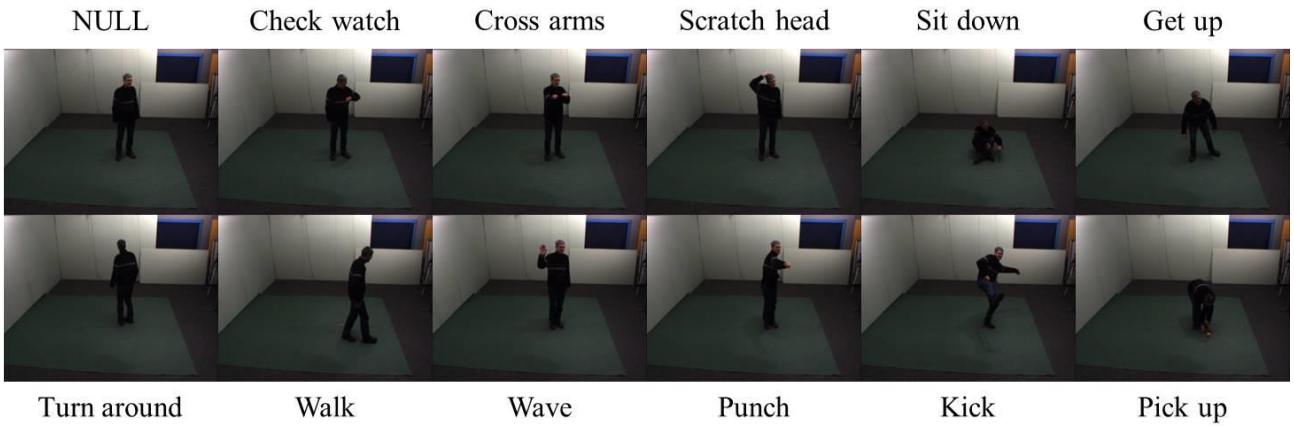

Figure 4. Examples of 12 used actions.

The IXMAS action dataset is a multi-view dataset for human action recognition. This dataset contains 15 different actions, 12 of which are used in this paper, and examples of these actions are shown in Figure 4. Specially, class 0 ``NULL", or called ``nothing", means no action is inside. Each of these actions is performed 3 times by 12 actors with 5 cameras in 23 fps. Following [29,30], 12 actions performed by 10 actors on 5 cameras are used to test our framework.

We adopt the leave-one-out (LOO) test strategy here. The IDT features of these videos are extracted in the first layer, and then these videos are divided into clips by Main Directions Initial Segmentation Method. It is worth noting that different

amounts of main directions can lead to great difference in quantity, length, and boundaries of these clips. Using a lot of directions may break a video to very short clips, while a small quantity of directions would make it insensitive to motion changes and would result in relatively long clips. Thus, three main directions are used in our implementation. Another important parameter is the hard threshold. The value of threshold is chosen by experiments and is set to be 6 in this paper.

4.2. Experimental results

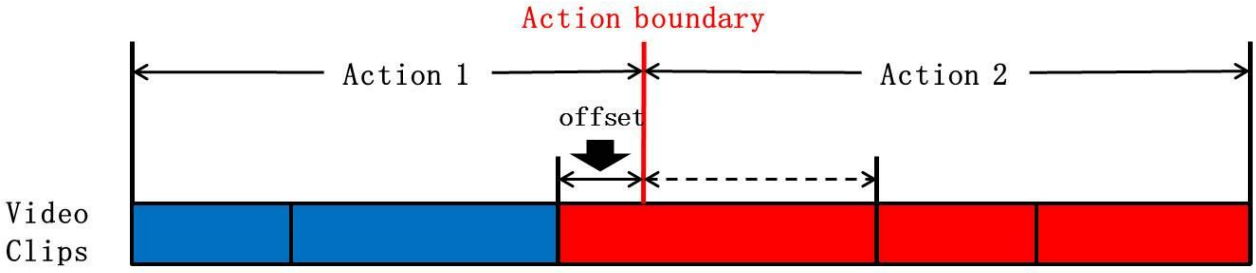

Figure 5. The action boundary offset. The input video is initially divided into 5 video clips. The action segmentation and recognition model annotates the first two clips as action 1 and classifies the following three clips to action 2. The red line indicates the actual boundary between the two actions. So, the action boundary offset is the distance between the end of the second clip and the actual boundary.

The concept of the action boundary offset is shown in Figure 5. It can be seen that one of the video clips contains frames of both actions, including the action boundary. The shortest distance between the boundaries of that clip and the action boundary, shown with solid line rather than dotted line, is the offset of this action boundary.

We compare the offsets of action boundaries between Main Direction Initial Segmentation Method and the method in [11] at frame level. Take videos recorded by camera 0 as example, there are 181 action boundaries in total, and all these boundaries are used to evaluate the distribution of offsets. The method in [11] divides an input video into fixed-length clips, and Figure 6 depicts offsets of this method with different lengths of video clips. Offsets of our method are shown in Figure 7, and the maximum length of video clips is set to be 15, 20, 25, and 30 frames. Figure 8 and Figure 9 compare the offset distribution and mean offset of the two methods. We can see that the offsets of the method in [11] are almost uniformly distributed, while our method is able to obtain much fewer offsets. This is because our method detects the motion boundaries according to the main direction of velocities and performs initial segmentation more reasonably. Comprehensively, our Main Direction Initial Segmentation Method can effectively reduce the amount of the error recognition caused by the fixed-length clips.

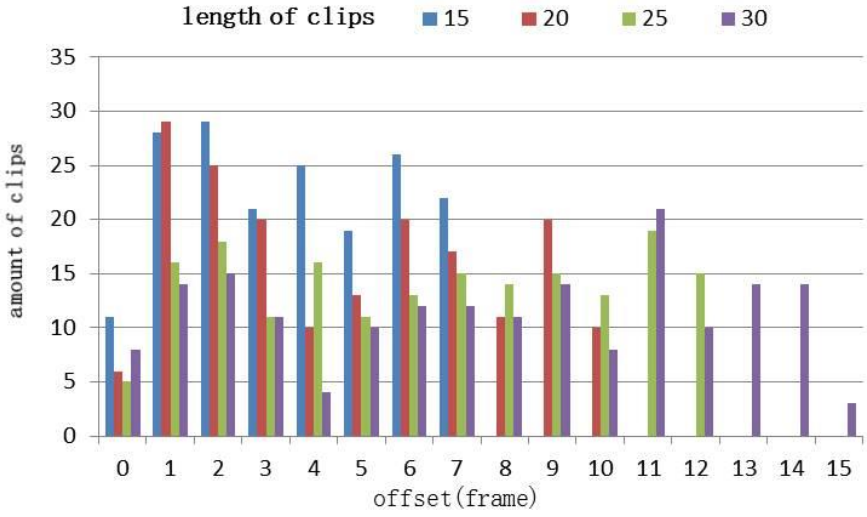

Figure 6. Offsets of the method in [11] with different clip lengths.

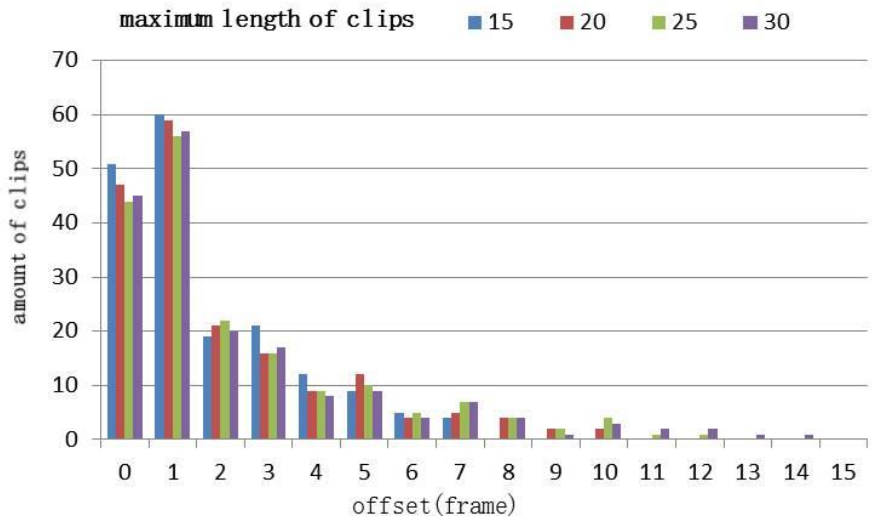

Figure 7. Offsets of Main Direction Initial Segmentation Method.

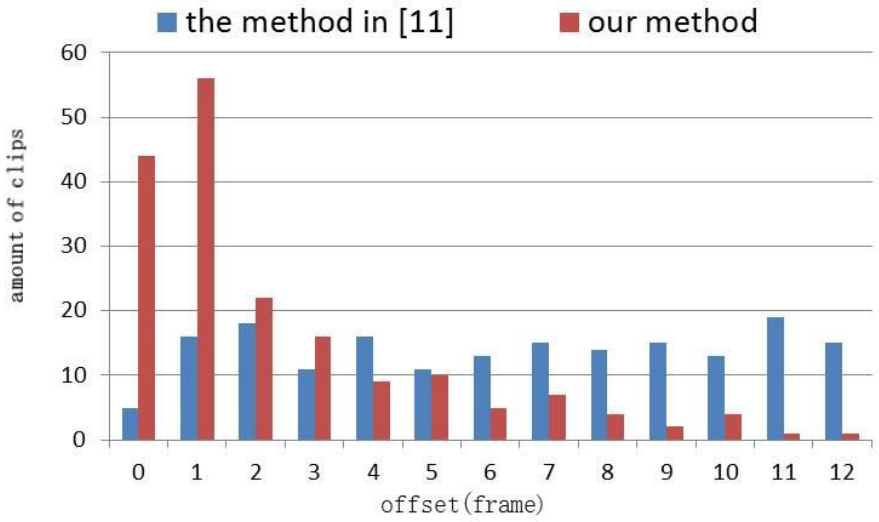

Figure 8. The comparison of the offset distribution between Main Direction Initial Segmentation Method and the fixed-length method in [11].

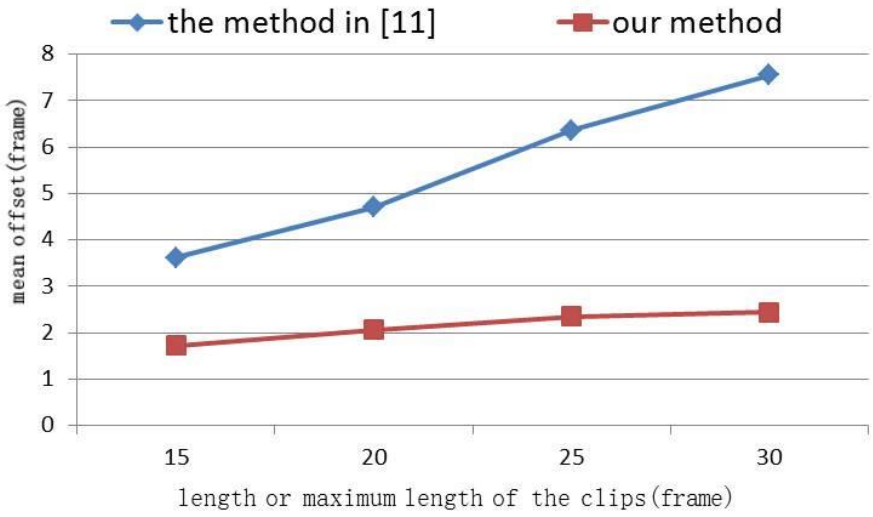

Figure 9. The comparison of the mean offset between Main Direction Initial Segmentation Method and the fixed-length method in [11].

We also compare the performance of action segmentation and recognition between the method in [11] and our double-layer framework. A 200-word dictionary of the trajectories is constructed by using k-means, and these words are employed to create representations of video clips. Then, the recognition method is executed on these video clips. The experimental results are illustrated in Figure 10, and Table 1 shows the numerical comparisons including the mean correct recognition rate. As is shown in Figure 10 and Table 1, our method outperforms the method in [11] for all of the 5 cameras, because the better initial segmentation results obtained by the Main Direction Initial Segmentation Method in the first layer can promote the performance of action segmentation and recognition in the second layer. The experimental results demonstrate that our double-layer framework is effective for action segmentation and recognition.

Table 1. The action assignment accuracy(%) of the method in [11] and our framework in frame level.

	camera 0	camera 1	camera 2	camera 3	camera 4	mean
The Method in [11]	61.41	54.70	57.33	61.57	49.73	56.95
Our Framework	77.03	70.23	58.95	65.89	50.54	64.53

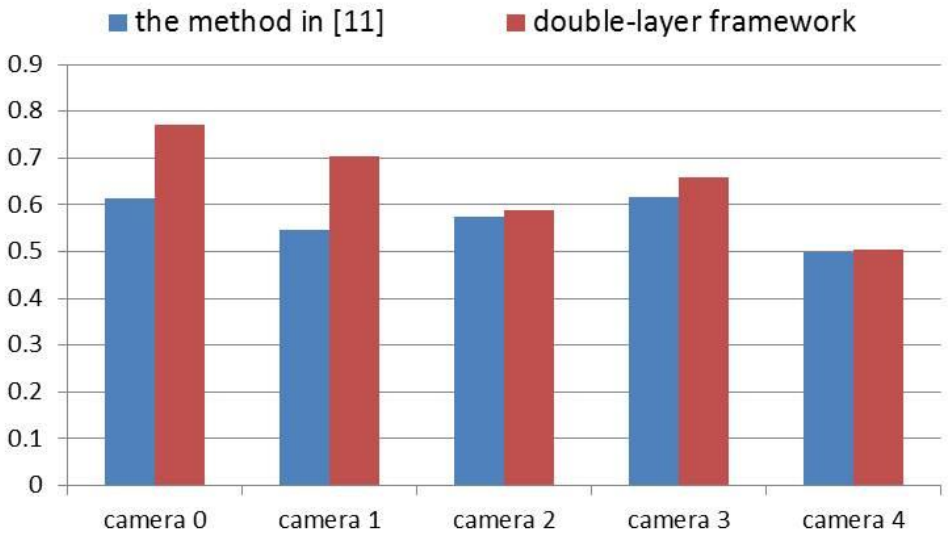

Figure 10. The comparison of action assignment accuracy between the method in [11] and our framework.

5. **Conclusions**

In this paper, we have presented a novel double-layer framework for joint segmentation and recognition of multiple actions in long-term videos. In the first layer, Main Direction Initial Segmentation Method is proposed to divide an input video into a series of clips. The second layer takes video clips as input, and adopts a learning method to group these clips into segments and annotate the action category of each segment simultaneously. Experimental results on the IXMAS action dataset have shown the effectiveness of the Main Direction Initial Segmentation Method and have proven that the proposed approach can achieve joint segmentation and recognition of multiple actions in videos. Recently, deep learning technology has proven its effectiveness in extracting discriminative representations for image and video classification. In future work, we plan to fuse IDT features and deep learning features to describe video clips.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No.61602320 and No.61170185, Liaoning Doctoral Start-up Project under Grant No.201601172, Foundation of Liaoning Educational Committee under Grant No.L201607 and No.L2014070, and the Young Scholars Research Fund of SAU under Grant No.15YB37.

References

1. J. K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," *Computer Vision & Image Understanding*, vol.73, no. 3, pp. 428-440, 1999
2. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE International Conference on Computer Vision*, vol.29, no.12, pp.1395-1402, Beijing, China, Oct 2005
3. A. Briassouli, T. Vagia, and K. Ioannis, "Human motion analysis via statistical motion processing and sequential change detection," *EURASIP Journal on Image & Video Processing*, vol. 2009, no. 1, pp. 1-16, 2009
4. E. J. Y. C. Cahuina and G. Camara Chavez, "A new method for static video summarization using local descriptors and video temporal segmentation," *26th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 226-233, Arequipa, Peru, August 2013
5. Q. Chen, Y. Cai, L. Brown, A. Datta, Q. Fan, R. Feris, and et al., "Spatio-temporal fisher vector coding for surveillance event detection," Proceedings of the 21st ACM international conference on Multimedia, pp. 589-592, Barcelona, Catalonia, Spain, October 2013
6. Y. Cheng, Q. Fan, S. Pankanti, and A. Choudhary, "Temporal Sequence Modeling for Video Event Detection," *27th IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, June 2014
7. K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, no.2, pp. 265-292, 2002
8. N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *18th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886-893, San Diego, CA, USA, June 2005
9. N. Dalal, B. Triggs, and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance," *9th European Conference on Computer Vision*, pp. 428-441, Graz, Austria, May 2006
10. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65-72, Beijing, China, Oct 2005
11. M. Hoai, Z. Z. Lan, and F. D. L. Torre, "Joint segmentation and classification of human actions in video," *24th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3265-3272, Colorado Springs, Colorado, USA, June, 2011
12. M. J. Rubin and W. A. Richards, "Boundaries of Visual Motion," *AI Memos*, vol. 835, 1985
13. J. Lei, G. Li, J. Zhang, Q. Gou, and D. Tu, "*Continuous action segmentation and recognition using hybrid convolutional neural network-hidden Markov model model*," Iet Computer Vision, vol.10, no.6 , pp.537-544, 2016
14. S. Li, K. Li, and Y. Fu, "Temporal Subspace Clustering for Human Motion Segmentation," *IEEE International Conference on Computer Vision*, pp. 4453-4461, Santiago, Chile, December 2015
15. G. Lu, M. Kudo, and J. Toyama, "Temporal segmentation and assignment of successive actions in a long-term video," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1936-1944, 2013
16. F. Lv and R. Nevatia, "Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching," *20th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, Minneapolis, Minnesota, USA, June 2007
17. D. Marr and L. Vaina, "Representation and recognition of the movements of shapes," *Proceedings of the Royal Society of London, Series B, Biological Sciences*, vol. 214, pp. 501-524, 1982
18. M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," *22th IEEE Conference on Computer Vision and Pattern Recognition*, pp.2929-2936, Miami, Florida, USA, June 2009
19. J. C. Niebles, C. W. Chen, and F. F. Li, "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification," *11th European Conference on Computer Vision*, pp.392-405, Heraklion, Crete, Greece, September 2010
20. A. S. Ogale, A. Karapurkar, G. Guerra-Filho, and Y. Aloimonos, "View invariant identification of pose sequences for action recognition," *VACE*, 2004
21. R. Polana and R. Nelson, "Low level recognition of human motion (or how to get your man without finding his body parts)," *Proceedings of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 77-82, Austin, Texas, USA, 1994
22. Y. Rui and P. Anandan, "Segmenting Visual Actions Based on Spatio-Temporal Motion Patterns," *IEEE Conference on Computer Vision and Pattern Recognition*, pp.111-118, Hilton Head, SC, USA, June 2000
23. C. Sch, I. Lapte, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *International Conference on Pattern Recognition*, vol.3, no.17, pp.32-36, Cambridge, UK, Aug 2004
24. L. Shao, L. Ji, Y. Liu, and J. Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recognition Letters*, vol.33, no.4, pp. 438-445, 2012
25. T. Syeda-Mahmood, "Segmenting actions in velocity curve space," *16th International Conference on Pattern Recognition*, pp. 1936-1944, Quebec, Canada, August 2002
26. K. Tang, "Learning latent temporal structure for complex event detection," *25th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1250-1257, Providence, Rhode Island, USA, June 2012
27. S. N. Vitaladevuni, V. Kellokumpu, and L. S. Davis, "Action Recognition Using Ballistic Dynamics," *21th IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-8, Anchorage, Alaska, USA, June 2008
28. A. Vögele and R. Klein, "Efficient unsupervised temporal segmentation of human motion," *Proceedings of the 2014 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 167-176, Copenhagen, Denmark, July 2014
29. D. Weinland, E. Boyer, and R. Ronfard, "Action Recognition from Arbitrary Views using 3D Exemplars," *IEEE International Conference on Computer Vision*, pp.1-7, Rio de Janeiro, Brazil, October 2007
30. D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision & Image Understanding*, vol.104, no.2, pp.249-257, 2006
31. H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," *IEEE International Conference on Computer Vision:*

3-6 December 2013; Sydney, Australia, pp. 3551-3558, 2013
32. J. Wang, X. Nie, Y. Xia, Y. Wu, and S. C. Zhu, "Cross-view action modeling, learning and recognition," *27th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2649-2656, Columbus, Ohio, USA, June 2014

Cuiwei Liu received her B.S. degree and Ph.D. degree from the Beijing Institute of Technology in 2009 and 2015, respectively. She is currently a lecturer with Shenyang Aerospace University, China. Her research interests include computer vision, machine learning, and video content analysis.

Yaguang Lu is a master student from the School of Information, Liaoning University. His research interests include machine learning, computer vision, and artificial intelligence.

Xiangbin Shi is Professor of Computer Science at Shenyang Aerospace University. He received the B.S. degree in Computer Application from the Shenyang University of Technology in 1985, the M.S. degree in Computer Application from the Northeastern University in 1990, and the Ph.D. degree in Computer Software and Theory from the Northeastern University in 1998. His current research interests include computer vision, virtual reality, and intelligent systems.

Deyuan Zhang received his B.S. degree and Ph.D. degree from Harbin Institute of Technology in 2002 and 2011, respectively. He is currently a lecturer with Shenyang Aerospace University, China. His research interests include computer vision, machine learning, and remote sensing image analysis.

Fang Liu received her B.S. degree and Master degree from LiaoNing University in 2004 and 2007, respectively. She is a PhD candidate in Northesatern University. She is currently a lecturer with Shenyang Aerospace University, China. Her research interests include computer vision, video understanding, and action Recognition.