A Violence Detection Approach Based on Spatio-Temporal Hypergraph Transition

Jingjia Huang, Ge Li *, Nannan Li, Ronggang Wang, and Wenmin Wang

School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055
jjhuang@pku.edu.cn, geli@ece.pku.edu.cn, linn@pkusz.edu.cn, rgwang@pkusz.edu.cn, wangwm@ece.pku.edu.cn

Abstract. In the field of activity recognition, violence detection is one of the most challenging tasks due to the variety of action patterns and the lack of training data. In the last decade, the performance is getting improved by applying local spatio-temporal features. However, geometric relationships and transition processes of these features have not been fully utilized. In this paper, we propose a novel framework based on spatio-temporal hypergraph transition. First, we utilize hypergraphs to represent the geometric relationships among spatia-temporal features in a single frame. Then, we apply a new descriptor called Histogram of Velocity Change (HVC), which characterizes motion changing intensity, to model hypergraph transitions among consecutive frames. Finally, we adopt Hidden Markov Models (HMMs) with the hypergraphs and the descriptors to detect and localize violence in video frames. Experiment results on BEHAVE dataset and UT-Interaction dataset show that the proposed framework outperforms the existing methods.

Keywords: Violence Detection, Action Recognition, Hypergraph, Spatio-temporal Feature, HMM

1 Introduction

Violence detection is one of the most challenging works in video processing. Considerable efforts have been done for its essential applications in video surveillance and smart camera systems. Since violence is a kind of aggressive interaction among multiple people with an unstable movement patterns, the major challenge for violence detection is to distinguish violent events from other normal human group activities. Therefore, in literature, violence detection has been considered to be a problem of abnormal detection [11, 12, 22] or group activity recognition[13, 14].

In order to automatically characterize violence in videos, some impressive works have been done. Most of these methods are based on spatio-temporal features [11, 12, 22, 15, 16] of human objects. Spatio-temporal features are always extracted from the areas called spatial temporal interest points (STIP),

proposed by Laptev and Lindeberg [17]. Spatio-temporal features are able to efficiently retain the information of motion variation in spatio-temporal domain. However, few researches concern a full expression of the relationships among the detected features. More specifically, there is a lack of expression on geometric relationships as well as the transition of these features. Geometric relationship is the representation of a posture in a single frame. Transition description is the representation of the posture variation process in consecutive frames. We then propose a framework based on spatio-temporal hypergraph transition to improve the performance of violence detection. On one hand, we propose an effective hypergraph model which can represent the concrete geometric relationships of the spatio-temporal features. On the other hand, we propose a descriptor called Histogram of Velocity Changing (HVC) to express the transition process among consecutive hypergraphs. HVC can efficiently distinguish normal actions such as walking, meeting and hugging ,which have a relatively stable as well as gentle movement patterns, from violence activities such as fighting which change drastically. The major contributions of our work are listed in the following:

(1) Introduce an innovative hypergraph to model the intra-frame relationships of local spatio-temporal features.

(2) Propose a novel descriptor called HVC to model the transition of spatio-temporal hypergraphs in consecutive frames for violence detection.

(3) Adopt a new spatio-temporal hypergraph transition based framework, which concatenates the hypergraph model and HVC to a Hypergraph-Transition Chain (H-T Chain) based on a redundant trajectory extraction algorithm.

Experiment results on UT-Interaction dataset and BEHAVE dataset demonstrate that the performance of our method outperforms the existing methods.

2 Related Work

Action recognition has been extensively studied in the last decade[5, 6, 1, 7]. However, violence detection is still a challenge problem due to the variety of patterns and the lack of training data. According to different application environments, various approaches are developed to solve the problem.

On the purpose of violence detection in movies, most of the works in the literature use audio features as an additional resource to represent the video elements [18]. [19] is one of the first proposals to characterize and index violent scenes in general TV drama and movies, in which characteristic sounds of violent events were used.

However, in surveillance systems, which is a more general application, people encounter the problems of no audio signal, low video resolution and relatively small scale of objects in a vast majority of cases. Therefore, many researchers have made efforts to figure out more suitable approaches to handle the problems. Hassner *et al.* propose a low-level descriptor for violence detection in crowded scenes called Violence Flow based on the magnitude changes of optical flow over successive frames [20]. Based on socio-psychological studies, Helbing *et al.* in [21]

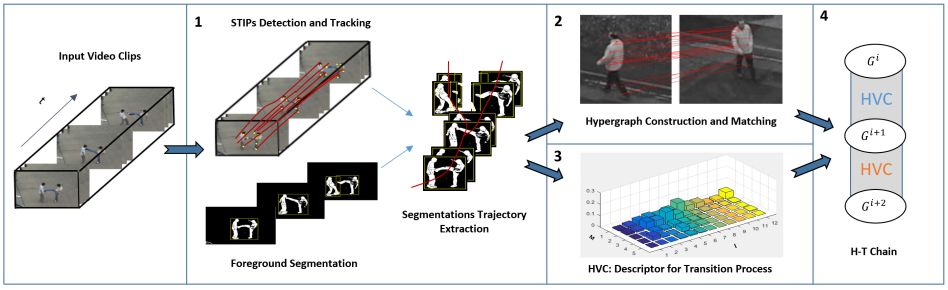


Fig. 1. A review for our framework. **Step 1**: pre-processing and trajectory extraction. **Step 2**: hypergraph model construction. **Step 3**: HVC descriptor construction. **Step 4**: H-T Chain construction based on the trajectory extracted in step 1. HMM is used to deal with the chains.

originally introduce a Social Force model to investigate the pedestrian movement dynamics. In the model, they treat the moving particles as individuals. How the individuals react to energy potentials caused by others and static obstacles through a repulsive force, is considered as an important clue for crowds activity recognition. Inspired by this work, Mehran *et al.* obtain Force Flow from Social Force Model and apply it to violence detection [12]. Based on STIP[17] trajectory, Cui *et al.* evaluate the anomaly in crowd with an interaction energy potential function [11]. Similarly to Cui *et al.*'s work, the approach proposed in [22] consider trajectories too. Hossein *et al.* develop a spatio-temporal descriptor called Histograms of Oriented Tracklets, which simultaneously captures the magnitude and orientation information of a set of STIP trajectories passing through a spatio-temporal volume.

Ryoo and Aggarwal [13] have demonstrated that the similarity between two videos can be well reflected by the structural similarity between sets of features extracted from them. On the other hand, graph is an effective tool for modeling complex structured data[16]. Consequently, some researchers adopt graphs to represent the structure of features in their works[15, 16, 23–25]. Brende and Sinisa propose a volumetric-based approach to learn spatiotemporal graphs of activities from videos [23]. In [15], Wu *et al.* construct two directed and attributed graphs based on intra-frame relationships and inter-frame relationships of the local features. Similarly, Aoun *et al.* construct Spatial Frame Graphs and Temporal Video Graphs for action recognition in [16]. Yi and Lin [24] construct undirected spatio-temporal graphs for each detected instance and attribute the spatio-temporal interactive relationship to the edges. However, all of them [15, 16, 24] are second-order matching methods which are limited to the affinity matrix embedding pairwise relationships between feature points with unary information, where the pairwise relationships are rotation-invariant but not scale-invariant as well as affine-invariant. Comparing to them, hypergrph based methods are more expressive with better integration of geometric information [26]. Therefore, Ta *et al.* propose a model for recognizing and localizing human

activities through hypergraph matching in [25]. In this paper, we also apply hypergraphs in our work. However, Ta *et al's* work is aiming at the recognition of individual activities while ours is able to analysis interactions among multiple people. Furthermore, we not only apply the posture information represented by hypergraphs but also the transition process between consecutive hypergraphs. Hypergraphs and corresponding HVCs are arranged in a H-T Chain to simulate the motion sequences.

3 Methodology

We propose a new framework to model group interactions for violence detection based on spatio-temporal hypergraph transition. **Fig. 1** is an illustration to the framework. We first extract STIPs and track them. Meanwhile, we use a foreground segmentation algorithm to detect the moving objects and assign STIPs to respective segmentations. Then, we construct hypergraphs for the segmentations and extract the trajectories of them. Transition descriptors are calculated between consecutive hypergraphs in the trajectory. Finally, for each trajectory, a H-T Chain is constructed and Hidden Markov Model (HMM) is used to detect violent events.

3.1 Pre-processing: Coarse Trajectory Extraction

During violence detection, interactions of a group of people instead of the individuals are the major concern, such as a group of people in fighting. Thus, a coarse object detection and tracking pre-processing step are employed in our frame work firstly. This will provide two benefit: (1) It helps to maintain the geometric information of the group. (2) It helps to analysis the motion transition in the following steps. As a result, we develop a simple but robust object trajectory extraction algorithm based on STIPs tracking in the framework.

As described in [17], Space-Time Interest Point (STIP) is an extension of the Harris corner detection operator to spacetime which is able to efficiently retain the information of motion variation in spatio-temporal domain. We use Laptev's release of STIP code for detection task and KLT tracker to track interest points. Then, we use Gaussian Mixture Model to get a set of foreground segmentations in tth frame:

$$S_t = \{s_{t,i}\}(i = 1...n) \tag{1}$$

where n is the number of segmentations in tth frame and the $s_{t,i}$ is

$$s_{t,i} = \{p_{t,j}\}(j = 1...m) \tag{2}$$

where $p_{t,j}$ is the jth STIPs in the segmentation and m is the total number of STIPs in it.

Comparing to the methods based on overlapping spatial sectors or sliding windows which break the original spatial relationships of STIPs [20, 13, 22], our

method clusters the STIPs belonging to the same segmentation as one analysis unit to benefit the further hypergraph construction.

Then we match s_{t,i_1} with s_{t+1,i_2} to obtain a segmentation trajectory if they satisfy:

$$\psi(s_{t,i_1}, s_{t+1,i_2}) = \frac{\sum_{j_1,j_2} \phi(p_{t,j_1}, p_{t+1,j_2})}{m_1} \geq thd \tag{3}$$

where m_1 is the number of STIPs in s_{t,i_1} and thd is a probability threshold for objects matching. $\phi(p_{j_1}, p_{j_2})$ is equal to 1 if p_{t+1,j_2} is the result of KLT tracker for p_{t,j_1}, otherwise to 0. The segmentation trajectory is for the further analysis in **Sect.3.3**.

3.2 Constructing and Matching Hypergraph Model

A hypergraph is a generalization of a graph in which an edge can join any number of vertices. Hypergraph is initially introduced to computer vision area for its ability of encoding higher order geometric invariants such as scale and affine invariants. In this paper, we use hypergraph to represent the posture information for segmentations by modeling the intra-frame relationships of STIPs. We define a hypergraph for a segmentation s_i as:

$$G^i = (V^i, E^i, F^i) \tag{4}$$

where V^i is a set of vertices *i.e* STIPs belonging to G^i. E^i is a set of hyperedges correspond to a 3-tuple of vertices. F^i is a set of feature vectors correspond to each vertices.

We consider two segmentations s_1 and s_2, and assume that N_1 and N_2 are the number of STIPs tracked in s_1 and s_2, respectively. Then we defined the hypergraph model for them as G^1 and G^2. A matching between G^1 and G^2 is equivalent to looking for an $N_1 \times N_2$ assignment matrix X such that $X_{i,j}$ is equal to 1 when p_i in G^1 matched to p_j in G^2, and to 0 otherwise. We follow the constrain in [27] that one node in G^1 can be matched to exactly one node in G^2 but no constrain to the nodes in G^2. Thus, we can get a set of assignment matrices:

$$A = \{X \in \{0,1\}^{N_1 \times N_2}, \Sigma_i X_{i,j} = 1\} \tag{5}$$

The measurement of similarity between two graphs can be formulated as the maximization of the following score on A :

$$score(A) = \sum_{i,i',j,j',k,k'} H_{i,i',j,j',k,k'} X_{i,i'} X_{j,j'} X_{k,k'} \tag{6}$$

where $H_{i,i',j,j',k,k'}$ is a similarity measure for the sets of features between hyperedges $E = \{i, j, k\}$ and $E^{'} = \{i^{'}, j^{'}, k^{'}\}$. The higher the value, the greater the similarity is. We utilize Duchenne *et al's* tensor-based power-iteration algorithm[27] for the optimization problem.

In order to adapt the hypergraph model to our work, we construct the similarity measurement $H_{i,i',j,j',k,k'}$ as follows. First, to measure the geometric similarity of two hypergraphs, we use the properties of the triangle formed by three STIPs. We denote a vector of the sine of the 3-tuple computed from their spatial coordinates as a_{ijk} :

$$a_{ijk} = [sin(\overrightarrow{ij}, \overrightarrow{ik}),\ sin(\overrightarrow{ji}, \overrightarrow{jk})] \tag{7}$$

Then, averaged optical flows, extracted from 3D patches around the corresponding STIPs, are used as motion descriptors which are denoted as F earlier in this section. As a result, we get the feature vector for a hyperedge E as:

$$f_{E_{ijk}} = [a_{ijk},\ F_i,\ F_j,\ F_k] \tag{8}$$

and calculate the similarity score with a Gaussian kernel:

$$H_{i,i',j,j',k,k'} = exp\{-\frac{\|f_{E_{ijk}} - f_{E_{i'j'k'}}\|_{l_2}}{\sigma}\} \tag{9}$$

where σ is the parameter of the kernel which governs the intra class variations of the features.

For the further analysis in the following section, we construct a codebook for the hypergraphs based on the proposed hypergraph similarity measurement. Spectral cluster is used for the unsupervised clustering.

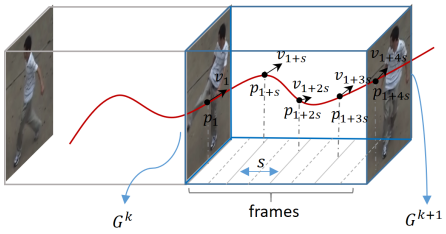
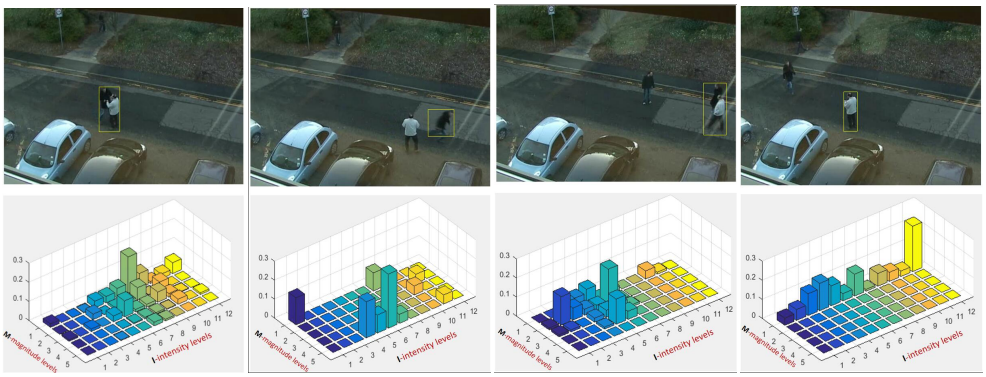


Fig. 2. Formation process for HVC. HVC for the transition from G^k to G^{k+1} is extracted from the portion of STIPs trajectory (red curve) passing through them. ν_i is the optical flow on the STIP p_i and s is the step length for velocity sampling.

3.3 Transition Description

If we consider the intra-frame relationships of local spatio-temporal features modeled with hypergraph as the postures in a motion sequence, then the transition of hypergraphs refer to the changing process from one posture to the next one. We produce Histogram of HVC for the transition description. A formation process is shown in **Fig. 2**. In **Sect.3.1** we extract trajectories of segmentations in which multiple STIP trajectories existed. The HVC computation for



(a) fighting (b) running (c) walking together (d) standing

Fig. 3. HVC examples for fighting, running, walking together and standing instances from BEHAVE dataset.

the transition from G^k to G^{k+1} starts with the portions of STIP trajectories passing through these two graphs. We denote this portion of the trajectory as $tr = \{p_1, ...p_i, ...p_n\}$, where i indicates the ith point in the STIP trajectory and n is the number of frames between G^k and G^{k+1}. Then, we define the intensity of velocity changing from p_i to p_{i+s} as:

$$I_{i,i+s} = \frac{\|\nu_i \ - \ \nu_{i+s}\|_{l_2}}{\|\nu_i\|_{l_2}} \tag{10}$$

where ν_i is the optical flow of p_i and s is the step length of sampling. Considering to get a more smooth intensity changing process, the sampling range can be expanded to G^{k-1} and G^{k+2} dependently. Besides, we compute the average velocity magnitudes of the sub-trajectory $\overrightarrow{p_i p_{i+s}}$ as:

$$M_{i,i+s} = \frac{\Sigma_{t=i}^{i+s} \|\nu_t\|_{l_2}}{s+1} \tag{11}$$

Finally, the intensities and magnitudes of all the sub-trajectories involving in the trajectories passing through G^k and G^{k+1} are quantized in I intensities and M magnitudes bins, respectively. Each of the sub-trajectories gives a contribution to the HVC histogram. Some HVC examples for different activities from BEHAVE dataset are shown in **Fig. 3**.

3.4 Violence Detection in Videos

In order to locate the violence in video clips, we use a sliding observation window to traverse the whole video. For segmentation trajectories within the window, a HMM is used as a classifier to determine whether a violence exist or not. A HMM sequence illustrated by **Fig. 4.(2)** is constructed with hypergraphs

and corresponding HVCs extracted from segmentation trajectories within the observation window shown in **Fig. 4.(1)**. We train a set of HMM models for different events including one for violence and the others for normal events. Given a behavior, we calculate probabilities of the HMM models, and we declare the behavior as a violence if the violence HMM model has the highest probability among all the HMMs.

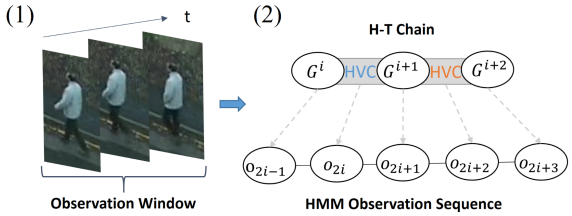


Fig. 4. (1) a motion sequence in an observation window. In (2), the top layer of graph is the H-T Chain for the motion sequence where G^i is the ith hypergraph in it. The second layer in (2) is the observation sequence for HMMs where o_{2i} is the $2i$th observation in it. The arrows, which connect two layers, indicate the correspondence between elements in the H-T Chain and observations for HMMs.

4 Experiments and Analysis

4.1 The Behave Dataset

Dataset. BEHAVE is a video dataset for multi-person behaviour analysis. There are a set of complex group activities annotated in the video, including meeting, splitting up, loitering, walking together, escaping, fighting and other interaction behaviours.

Parameters. We fixed the $thd = 0.2$ for trajectory extraction in Equation 3. For the hypergraphs, spectral cluster is used and a codebook with 25 vocabularies is constructed. For the HVC, we fixed $M = 5$ and $I = 12$. Meanwhile, a codebook with 15 vocabularies is trained for HVC. To construct a H-T Chain for a motion sequence, we set the size for observation window as 80 frames while the step length is 20 frames. For the portion of segmentation trajectory within the window, the interval size between two hypergraphs is set to 3 frames and the sampling range for the HVC is 9 frames with a step length of 3 frames.

Evaluation. The detection task of BEHAVE dataset is a violence/non-violence detection. The comparison methods include the optical flow based method, social force model[12] and interaction energy potential [11]. Following settings in [11], we used half of normal and abnormal videos for training and the rest for testing. According to the past practice, the results are evaluated by the means of ROC as shown in **Fig.5**.

Analysis. According to the ROC curves in **Fig.5**, our method outperforms the Interaction Energy Potentials[11], Social Force Model[12] and Optical Flow,

which demonstrates that our framework is competitive with the existing methods on BEHAVE dataset.

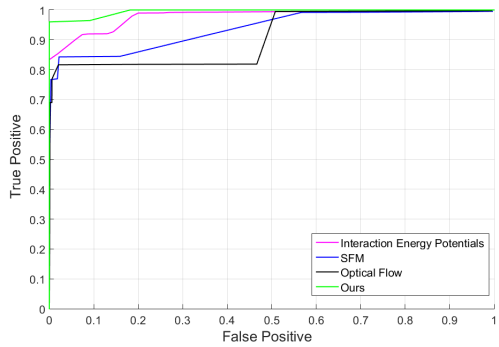


Fig. 5. ROC curves on BEHAVE dataset. Our method (green line) is compared with Interaction Energy Potentials[11] , Social Force Model[12] and Optical Flow.

4.2 The UT-Interaction Dataset

Dataset. The UT-Interaction dataset includes two sets. UTI$^{\sharp 1}$ was captured on a parking lot with mostly static background, while UTI$^{\sharp 2}$ was captured on a lawn with slight background dynamics and camera jitters. Both the sets contain 60 videos of 6 categories of human interactions: push, kick ,punch, shake hands, point and hug(10 videos for each category).
Parameters. We keep the same settings for UTI$^{\sharp 1}$ and UTI$^{\sharp 2}$. We fixed the $thd = 0.3$ for trajectory extraction in Equation 3 and we set the size of observation window as the length of each clips since the duration of videos is short. The other parameters are as same as that in the experiment for BEHAVE.
Evaluation. We evaluate the performance of our framework via the leave-one-out cross validation strategy proposed in [13]. The confusion matrixes of our recognition results on UTI$^{\sharp 1}$ and UTI$^{\sharp 2}$ are shown in **Fig. 6.(a)** and **Fig. 6.(b)**. Besides, since our method address on the problem of violence detection, we also perform the violence/non-violence recognition experiments on this dataset. First, we divide the action categories into two groups, where push, kick and punch are labeled as violent actions; while shake hands, point and hug are labeled as non-violent actions. Such classification is shown with different colors (gray for violent group and blue for non-violent group) in **Fig. 6**. Then, for each group, we choose the maximum log probability of the three actions' HMMs as the score for the entire group. For instance, if a clip to recognize has the log probabilities of HMMs for kick, push and punch as -0.43, -1.30, and -2.71, the score of the violent group is selected as -0.43 (the maximum one). Finally, we classify the actions as violence or non-violence via the normalized group scores. **Table 1** shows our experiment results comparing to the approaches of Ryoo & Aggarwal[13], [8], Ryoo[10], Ke[4], Yu & Yun[3] and Xu *et al.*[9]. In the table,

(a) UTI$^{\sharp 1}$ (b) UTI$^{\sharp 2}$

Fig. 6. Confusion matrix of our method on UT-Interaction dataset

Table 1. Experiments results on UT-Interaction dataset. **'-'** means no corresponding experiment results.

Methods	Accuracy$^{\sharp 1}$-α	Accuracy$^{\sharp 1}$-β	Accuracy$^{\sharp 2}$-α	Accuracy$^{\sharp 2}$-β
Ryoo & Aggarwal[13]	70.80%	-	-	-
MSSC[8]	83.33%	-	81.67%	-
Ryoo[10]	85.00%	-	70.00%	-
Ke[4]	93.33%	-	-	-
Yu & Yun[3]	93.33%	96.67%	**91.67%**	95.00%
Xu *et al.*[9]	**96.67%**	96.67%	90.00%	-
Ours	**96.67%**	**98.33%**	90.00%	**100.00%**

Accuracy$^{\sharp 1}$-α and Accuracy$^{\sharp 2}$-α mean the recognition accuracy for the six categories of actions in UTI$^{\sharp 1}$ and UTI$^{\sharp 2}$; Accuracy$^{\sharp 1}$-β and Accuracy$^{\sharp 2}$-β are the violence/non-violence recognition accuracy in the two sets.

Analysis. On the dataset, [13] is the benchmark for action recognition task. As it illustrated in column 1 and 3 of **Table 1**, for action recognition task, our method achieves the same performance as state-of-the-art methods[9] on UTI$^{\sharp 1}$ and has a competitive performance on UTI$^{\sharp 2}$; while for the violence/non-violence recognition task, which are presented in column 2 and 4, we outperform state-of-the-art methods on both sets. We notice that though [3] has a better performance than ours on UTI$^{\sharp 2}$, we get a higher accuracy for violence recognition. It is because that [3] is confused by hug and punch that belong to non-violence and violence group respectively. Thanks to HVC, which is sensitive to the motion changing intensity, our method is able to distinguish the gentle movement from vigorous movement. Therefore, less confusion exists between non-violent and violent actions in our methods. The results demonstrate that our method achieves superior performance over these comparison methods.

5 Conclusion

We proposed a spatio-temporal hypergraph transition based framework for violence detection. In our method, We introduced an innovative hypergraphs to

model the intra-frame relationships of local spatio-temporal features. Besides, a novel descriptor called HVC is proposed to model the velocity changing intensity during the transition process of hypergraphs among consecutive frames. A motion sequence is represented with a H-T Chain, which was constructed with hypergraphs and HVCs. HMM was used as a classifier in the framework to indicate violent/nonviolent event. Experiment results on UT-Interaction dataset and BEHAVE dataset demonstrated the superiority of our method.

Acknowledgments. This work was supported by Guangdong Province Projects of 2014B010117007, National Science Foundation of China (No.U1611461), National Natural Science Foundation of China(61602014), Shenzhen Peacock Plan (20130408-183003656), and Science and Technology Planning Project of Guangdong Province, China (No. 2014B090910001).

References

1. Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," *2016 European Conference on Computer Vision*, pp. 20–36, 2016.
2. M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2650–2657.
3. Kong Yu and Fu Yun, "Close human interaction recognition using patch-aware models," in *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 2015, vol. 25, pp. 167–178.
4. Qiuhong Ke, Mohammed Bennamoun, Senjian An, Farid Boussaid, and Ferdous Sohel, "Human interaction prediction using deep temporal features," in *2016 European Conference on Computer Vision*.
5. Tran Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," *IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.
6. Limin Wang, Yu Qiao, and Xiaoou Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4305–4314, 2015.
7. Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang, "Real-time action recognition with enhanced motion vector cnns," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
8. Tian Lan, Tsung Chuan Chen, and Silvio Savarese, *A Hierarchical Representation for Future Action Prediction*, Springer International Publishing, 2014.
9. Zhen Xu, Laiyun Qing, and Jun Miao, "Activity auto-completion: Predicting human activities from partial videos," in *International Conference on Computer Vision*, 2015, pp. 3191–3199.
10. M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 1036–1043.
11. Xinyi Cui, Qingshan Liu, Mingchen Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *The IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, Co, Usa, 20-25 June*, 2011, pp. 3161–3167.

12. R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 935–942.
13. M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *IEEE International Conference on Computer Vision*, 2009, pp. 1593–1600.
14. S. J. Blunsden and R. B. Fisher, "The behave video dataset: ground truthed video for multi-person," in *Annals of the Bmva*, 2009, vol. 4.
15. Baoxin Wu, Chunfeng Yuan, and Weiming Hu, "Human action recognition based on context-dependent graph kernels," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2609–2616.
16. Najib Ben Aoun, Mahmoud Mejdoub, and Chokri Ben Amar, "Graph-based approach for human action recognition using spatio-temporal features," *Journal of Visual Communication & Image Representation*, vol. 25, no. 2, pp. 329–338, 2014.
17. Ivan Laptev and Tony Lindeberg, "On space-time interest points," 2005, vol. 64, pp. 107–123.
18. Fillipe D. M. De Souza, Guillermo C. Chavez, Eduardo A. Do Valle, and Arnaldo De A. Araujo, "Violence detection in video using spatio-temporal features.," in *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*, 2010, pp. 224–230.
19. Je Ho Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in *International Conference on Image Processing, 1998. ICIP 98. Proceedings*, 1998, pp. 353–357.
20. T Hassner, Y Itcher, and O Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–6.
21. Dirk Helbingp&Eacute and Ter Moln&Aacute, "Social force model for pedestrian dynamics," *Physical Review E Statistical Physics Plasmas Fluids & Related Interdisciplinary Topics*, vol. 51, no. 5, pp. 4282–4286, 1995.
22. Hossein Mousavi, Hamed Kiani Galoogahi, Alessandro Perina, and Vittorio Murino, *Detecting Abnormal Behavioral Patterns in Crowd Scenarios*, Springer International Publishing, 2016.
23. William Brendel and Sinisa Todorovic, "Learning spatiotemporal graphs of human activities.," in *IEEE International Conference on Computer Vision*, 2011, pp. 778–785.
24. Yang Yi and Maoqing Lin, "Human action recognition with graph-based multiple-instance learning," *Pattern Recognition*, vol. 53, no. C, pp. 148–162, 2016.
25. Anh Phuong Ta, Christian Wolf, Guillaume Lavou, and Atilla Baskurt, "Recognizing and localizing individual activities through graph matching," in *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 196–203.
26. Soonyong Park, Sungkee Park, and Martial Hebert, "Fast and scalable approximate spectral matching for higher order graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 479–492, 2014.
27. O Duchenne, F Bach, In So Kweon, and J Ponce, "A tensor-based algorithm for high-order graph matching.," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no. 12, pp. 2383–95, 2011.