**Paper:**

# A Skeleton-based Method for Recognizing the Campus Violence

**Yi Xing, Yaping Dai, Kaoru Hirota, Zhiyang Jia***

School of Automation, Beijing Institute of Technology
No.5 Zhongguancun South Street, Haidian District, Beijing 100081, P.R.China
E-mail: yi.xing@bit.edu.cn

**Abstract. With the surveillance video acquired by campus surveillance devices as input, this paper adopts an action recognition method to determine whether or not campus violence has occurred. We incorporate attention module into the two-stream adaptive graph convolutional network. This method can process skeletal information extracted from the video with graph convolution network. And then it determines the presence of school violence through action recognition based on the human skeleton. To verify the validity of this method, we do experiments on the filtered NTU RGB+D dataset. The accuracy of our method is 96.07%, which is 1.31% higher than the conventional method.**

**Keywords:** Campus violence; Attention module; Graph convolutional network; Skeletal information.

## 1. INTRODUCTION

In this fast-changing world, more and more adolescents have seriously affected by the violent contents from the Internet. It also leads to an increasing incidence of campus violence. Based on the campus violent actions' complexity and rapidity, it is necessary to use computer vision methods with school surveillance video to monitor campus violence. Some researchers collect data by wearable devices, in order to detect violent actions with expression, voice and body. With the development of video recording devices and networks, it is more important to analyze and understand human actions from video information, the optical-flow method is widely used in violent action recognition. Wang et al.[1] first proposed the dense trajectories(DT) algorithm. The DT algorithm samples feature points densely at multiple scales of the image by grid segmentation. Secondly, it calculates the number of optical-flow in the sampled feature field. Finally it can obtain the trajectory of the feature point. Subsequently, deep learning was adopted for action recognition, these end-to-end networks could automatically learn features from images and output classification results. Wang et al.[2] propose to learn actions based on RGB images and identify them with optical-flow field. The research method with wearable devices can not carry out contactless action recognition and is difficult to be used in detecting campus violence. The method based on image processing and optical-flow is difficult to extract features from background images. It is usually effectd by weather, lighting, camera angles and clothing, and it is more susceptible to background noise. At present, human skeletal information can be extracted from the image, and action recognition based on skeleton can be better applicable to dynamic environment or complex image background.

In order to avoid the above problems, in the research of violent action recognition, we extract skeletal information about the human body in the video based on human pose estimation. A deep learning approach with human skeleton information is adopted to identify violent actions at school. We adopt graph convolutional network(GCN) to recognize the violent actions. It is different from the traditional image processing with convolutional neural network, but rather recognize violence actions by the graph approach. This paper adopts two-stream attention in adaptive graph convolutional network(2s-AAGCN) to recognize the violent actions. (1)We adopt OpenPose to extract the motion skeleton of students in campus surveillance video, and then we convert the skeletal data from 2D to 3D for action recognition. (2)Based on spatial temporal graph convolutional network(ST-GCN) and two-stream adaptive graph convolutional network(2s-AGCN), we add attention module to form 2s-AAGCN, which can enhance the network's ability to extract spatial features. (3)Experiments on the filtered dataset are carried out and compared with other gcn method, which proves the effectiveness of our method.

In the remainder of this paper, we first provide some related work in 2. In 3, we introduce the method to identify violent actions, including the extraction of human skeletal information and the specification of 2s-AAGCN. In 4, we summarize and analyze the experimental results. Finally, we make conclusion and point out future research direction in 5.

## 2. RELATED WORK

Human action recognition is a complex problem, it can be divided into four levels according to the degree of complexity. (1) Gesture. It refers to the movement of various parts of the human body, such as "waving hands", "head down". (2) Action. It refers to the activities of a single person, such as "jumping", "runing". (3) Interaction.

It refers to the interaction between people and people or between people and things, such as “two people fighting”, “taking things”. (4) Group activity. It is an activity among many people, such as “a group of people in a parade”, “many people in a dance”. The action recognition complexity classification is shown in **Fig.1**.

(a) Gesture

(b) Action

(c) Interaction

(d) Group activity

**Fig. 1.** The classfication of human action recognition.

## 2.1. Traditional Skeleton-based Action Recognition

Some traditional methods require hand-crafted features and traversal rules to achieve skeleton-based action recognition. Shotton et al.[3] propose a fast and accurate approach to predict body joint 3D location from a single depth image that does not require temporal information. Based on this method. Xia et al.[4] use 3D joint node histograms to represent the body pose, and the body action is built model by the Discrete Hidden Markov Model. Their main components of the algorithm are real-time, including the extraction, calculation and classification to these 3D skeletal joint positions. Using the method of skeletal joint point trajectories allows the correspondence between joint points to be obtained from different angles. Pazhoumanddar et al.[5] use the longest common subsequence to extract more robust action features. Algorithms select high-resolution features from the relative motion trajectories of the skeleton to describe the relevant motion. However, the performance of these methods depends on the estimation of the human position. When some joint points are sheltered by other objects, the lost and incorrect parts may affect the results of action recognition. While Keceli et al.[6] extract human action features based on angular and displacement information from skeletal nodes, the action classification with support vector machines(SVM) and random forest algorithms. This method does not rely on the dataset, and it’s smaller in computation compared with those trajectory feature methods.

## 2.2. Deep Network Skeleton-based Action Recognition

Deep learning methods contain RNN-based methods, CNN-based methods and GCN-based methods. With the development of deep learning, RNN-based methods appear gradually. Du et al.[7] divide the human skeleton into five parts according to human physical structure, and then separately fed them to five bidirectionally recurrently connected subnets. The network is a good application that the recurrent neural network (RNN) can model temporal series contextual information well. Song et al.[8] propose an end-to-end spatial and temporal attention model. This model can learn to selectively focus on discriminative joints of skeleton within each frame of the inputs. So it can pay attention of different levels to the outputs in different frames. Zhang et al.[9] design a view adaptive RNN with LSTM architecture. It enables the network itself to adapt to the most suitable observation viewpoints from end to end. Liu et al.[10] propose a spatio-temporal based LSTM model. It can analyze the 3D position of the skeletal joints at each frame and each processing step, while use spatiotemporal features. In recent years, a number of CNN-based approaches have also emerged. Kim et al.[11] redesign the original TCN by factoring out the deeper layers into additive residual terms, which yields both interpretable hidden representations and model parameters. Liu et al.[12] propose an enhanced skeleton visualization method to represent a skeleton sequence as a series of visual and motion enhanced color images. It can implicitly describe spatiotemporal skeleton joints in a compact and distinctive manner. Li et al.[13] design a novel skeleton transformer module to rearrange and select important skeleton joints automatically. Li et al.[14] propose an end-to-end convolutional co-occurrence feature learning framework to aggregate different levels of contextual information.

The skeleton graph of the human body is a topological graph, CNN can not handle non Euclidean structure’s data. In the topological graph, the number of neighbors of each vertex may be different, so we can not use a convolution of the same size kernel to perform convolutional operations. Graph convolutional network(GCN) is based on the CNN, it can efficiently extract spatial features for machine learning in such data. Human action data is a sequence of time series with spatial features. How to comprehensively explore the spatio-temporal features of motion through graph convolutional networks is the current research hotspot in the field of behavior recognition. For the skeleton-based action recognition task, Yan et al.[15] first apply GCN to model the skeleton data. They construct a spatial graph based on the natural connections of joints in the human body and add the temporal edges between corresponding joints in consecutive frames. A distance-based sampling function is proposed for constructing the graph convolutional layer. It is employed as a basic module to build the final spatiotemporal graph convolutional network. Li et al.[16] design multiscale convolutional filters to encode the graph structure data and propose a recursive graph convolutional network.

## 3. METHODOLOGY

### 3.1. Construct Skeleton-graph by OpenPose

Action recognition methods based on skeletal data have been widely studied and valued for their adaptability to dynamic environments and background complexity. There are CPM[17], CPN[18] and OpenPose[19] for human body posture estimation. Among them, OpenPose is widely used because of faster process and higher accuracy to extract the skeleton.

OpenPose is an open source human posture recognition library that uses convolutional neural networks and supervised learning developed on caffe framework. OpenPose represents the first real-time multi-person system to jointly detect human body, hand, facial, and foot keypointswith excellent robustness. OpenPose uses a down-top approach, including parts detection and parts association. It can detect the location of each key point separately, then obtain the heatmaps for predicting each key point in the body. The Gaussian peak at each human key point represents the neural network believes a human key point in there. The keypoint detection results are linked to determine which person in the picture each keypoint specifically belongs to. The 2D skeleton sequence is extracted from the video by OpenPose, then we can get three sets of information about *x*, *y*, *acc*. (*x*,*y*) is the label of the keypoints, *acc* is the keypoint's accuracy. In order to transform 2D skeleton data into 3D skeleton, we can use the *acc* as the z-axis to obtain a 3D skeleton sequence to be the input data. In school violence detection, we can use the above method to extract 3D skeletons of students, and then it can be used for action recognition's input data to our action recognize model. The campus violence skeleton-graph by OpenPose is shown in **Fig.2** left, and the joints are labeled as shown in **Fig.2** right.

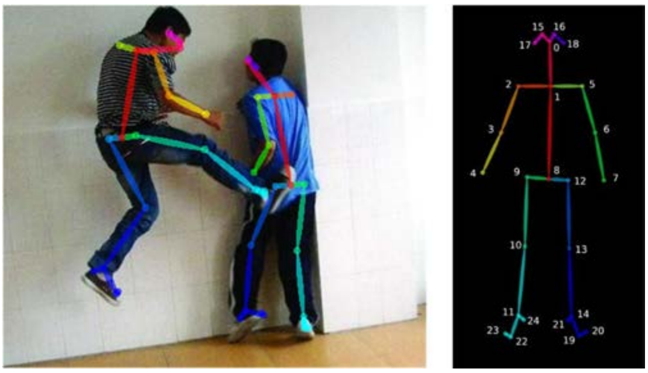

**Fig. 2.** Illustration of the campus violence skeleton-graph on the left. Illustration of the joints configuration by OpenPose on the right.

### 3.2. Two-stream Attention Adaptive Graph Convolutional Network

Yan et al.[15] take joints as nodes and the connections between nodes as edges to construct the skeleton graph. **Fig.3** shows an example of a spatial–temporal skeleton graph on the left.

Spatial graph convolution is formulated as:

$$f_{\text{out}}\left(v_i\right) = \sum_{v_j \in B_i} \frac{1}{Z_{ij}} f_{in}\left(v_j\right) \cdot \omega\left(l_i\left(v_j\right)\right) \tag{1}$$

Where $f$ is the feature map. $v_i$ is the vertex of the graph. $B_i$ denotes the sampling area of the convolution for $v_i$, which is defined as the one-distance neighbor vertexes($v_j$) of the target vertex($v_i$). The neighbor set $B_i$ of a joint node $v_i$ is partitioned into a fixed number of $K$ subsets, where each subset has a numeric label. The mapping function $l_i$ maps a node in the neighborhood to its subset label. $\omega$ is the weighting function similar to the original convolution operation, which gives different weights according to different $l_i$ values. The normalizing term $Z$ equals the cardinality of the corresponding subset. The authors of ST-GCN propose three partitioning strategies and illstrate the best strategy. The best spatial configuration partitioning strategy is shown in **Fig.3** right. So this paper directly adopts this strategy.

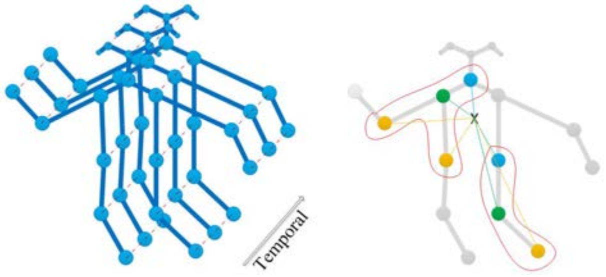

**Fig. 3.** Illustration of the spatiotemporal graph used in st-gcn on the left. Illustration of the mapping strategy on the right.(Different colors denote different subsets)

#### 3.2.1. Attention Adaptive Graph Convolutional Network

To implement the ST-GCN, [16]transformed Equation (1) into:

$$f_{\text{out}} = \sum_{k}^{K_v} \mathrm{W}_k \big(f_{in} \Lambda_k^{-\frac{1}{2}} \mathrm{A}_k \Lambda_k^{\frac{1}{2}}\big) \odot \mathrm{M}_k \tag{2}$$

$K_v$ denotes the kernel size of the spatial dimension. $W_k$ is the $C_{out} \times C_{in} \times 1 \times 1$ weight vector of the $1 \times 1$ convolution operation, which represents the weighting function $\omega$ in Equation(1). $M_k$ is an $N \times N$ map that indicates the importance of each vertex. $\odot$ denotes the dot product. [20] proposes the spatiotemporal graph convolution for the skeleton data described above is calculated based on a predefined graph, which may not be the best choice. According to Equation(2), the topology of the graph is actually decided by the adjacency matrix and the mask matrix, i.e. $A_k$ and $M_k$, [20] makes the graph structure adaptive by changing Equation(2) into the following form:

$$f_{\text{out}} = \sum_{k}^{K_v} \mathrm{W}_k \; f_{\text{in}}\left(\mathrm{A}_k + \mathrm{B}_k + \mathrm{C}_k\right) \tag{3}$$

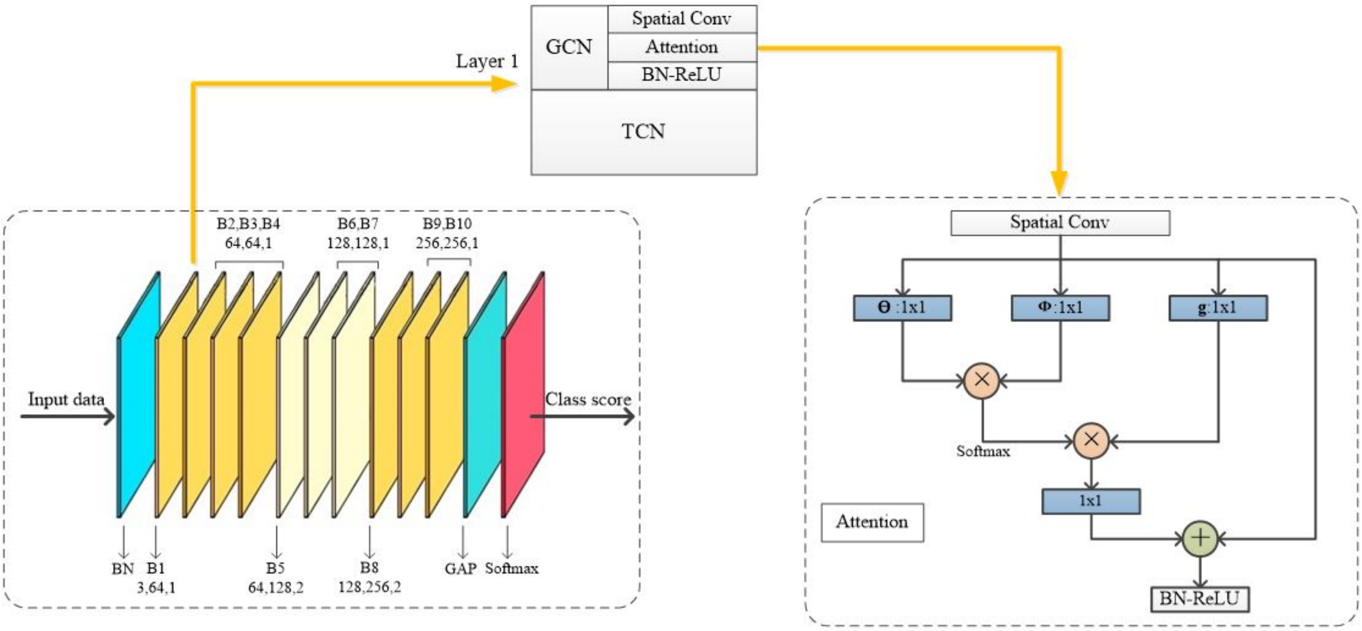

**Fig. 4.** Illustration the attention module in each AAGCN block.

$A_k$ is the same as the original normalized $N \times N$ adjacency matrix $A_k$ in Equation(2). It represents the physical structure of the human body. The elements of $B_k$ are parameterized and optimized together with the other parameters in the training process. There are no constraints on the value of $B_k$, which means that the graph is completely learned according to the training data. With this data-driven manner, the model can learn graphs that are fully targeted to the recognition task and more individualized for different information contained in different layers. $C_k$ is a data-dependent graph which learn a unique graph for each sample. [20] applies the normalized embedded Gaussia function to calculate the similarity of the two vertexes. The adaptive graph convolutional network replaces $A_k$, $M_k$ with $A_k$, $B_k$, $C_k$. In this way, it can strengthen the adaptive and flexibility of the model without degrading the original performance.

In the adaptive graph convolution network model, the receptive field of the convolution operation is the one-neighbor of the root node, so it only captures local features. However, in a different sample of different action classes, the relationship between the joints is not limited to the one-neighbor of the joint. The importance of different trunks is variant in human movement. For example, the movement of the neck may be less important than the legs, by which we can even judge running, walking and jumping, but in the movement of the neck may not contain much valid information. The attention module that we adopted is based on non-local neural networks [21]. It defines a generic non-local operation in deep neural networks. A non-local operation is a flexible building block and can be easily used together with convolutional or recurrent layers. It refines the information for a feature map, which is a good implementation of the attention module. Also, it can be added into the earlier part of deep neural networks, unlike FC layers that are often used in the end. This allows us to build a richer hierarchy that combines both non-local and local information.

We introduce non-local neural networks into AGCN that can focus on the features of all joints and get more efficient features. **Fig.4** shows the network structure of attention in adaptive graph convolutional network(AAGCN), where we add the attention module after the spatial convolution operation(Spatial Conv). Firstly, on the feature map of Spatial Conv we complete $1 \times 1$ convolution to get the $\theta$, $\phi$, $g$ features. Secondly, we perform a matrix point multiplication operation on $\theta$ and $\phi$ to calculate the autocorrelation in the feature. And then we implement Softmax operation to obtain the attention coefficient. Thirdly, we multiply the attention coefficient and the feature matrix $g$. Finally, residual connection is established with the original input feature map and then we get a new set of features. We add the above attention module in each AGCN block between Spatial Conv and the layer of BN-ReLU. Our network is composed by ten basic blocks as the above-mentioned. It is shown in **Fig.4**. The numbers of output channels for each block are 64, 64, 64, 64, 128, 128, 128, 256, 256 and 256. A data BN layer is added at the beginning to normalize the input data. A global average pooling layer is performed at the end to pool feature maps of different samples to the same size. The final output is sent to a softmax classifier to obtain the prediction.

### 3.2.2. Two-stream: Joint and Bone

Violence in schools occurs in two-player interactions, with traditional ST-GCN focusing on the roll-up of joints. For violent actions such as hitting and kicking, bone information and joint information are equally important. [20] defines that each bone is defined between two joints, so the bone is representedas a vector pointing to its target joint from its source joint, which contains not only the length information, but also the direction information. We adopt this strategy to train network model on these bones and joints points, then respectively obtain the J-AAGCN and the B-AAGCN. The two-stream about joint-stream and bone-stream is shown in **Fig.5**.

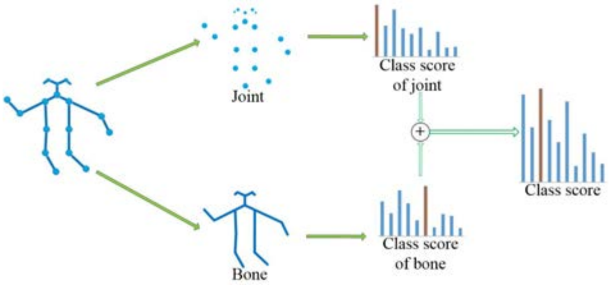

**Fig. 5.** Illstrate two-stream class score ensemble about joints and bones.

We respectively extract and process joints' information and bones' information, and finally we obtain the final prediction by the scores of two streams fused.

## 4. EXPERIMENTS

### 4.1. Dataset

In order to comparing our experiments with traditional graph convolutional networks used for violence recognition, our experiments are based on NTU RGB+D[22]. This dataset is a classical dataset about action recognition in recent years. NTU RGB+D is the largest and widest used action recognition dataset currently, which contains 56,880 action clips in 60 action classes. The dataset is captured by three Kinect V2 cameras concurrently. The resolutions of RGB videos are 1920 × 1080, depth maps and IR videos are all in 512 × 424, and 3D skeletal data contains the 3D coordinates of 25 body joints at each frame. These 56,880 action clips are performed by 40 volunteers in different age groups ranging from 10 to 35. This dataset provides 3D joint locations of each frame detected by Kinect depth sensors. There are 25 joints for each subject in the skeleton sequences, while each video has no more than 2 subjects. Most of the 60 action classes are single-person action in the NTU RGB+D dataset , in which we screen 10 classes about two-person interaction. Including: *punching*, *kicking*, *pushing*, *point finger*, *hugging*, *giving object*, *touch pocket*, *shaking hands*, *walking towards*, *walking apart*. The chosen classes include *pushing*, *punching* and *kicking*, which are common violence actions at school. And it can be used to train and verify effectiveness with the above method for violent action recognition.

### 4.2. Properties of Our Method

Our experimental environment is under ubuntu 18.04 with nvidia GTX 1080Ti, the algorithm is conducted on the Pytorch deep learning framework, the program code language is python.

We examine the effectiveness of the 2s-AAGCN with the filtered NTU RGB+D dataset. This paper focuses on the recognition of violent actions in campus surveillance, its application scenario is usually a certain perspective's action recognition for different students. So our experiments are implemented with X-subject setting. In this setting the training clips come from one subset of actors and the models are evaluated on clips from the remaining actors. We use the AAGCN model for training and validating on the preprocessed joint and bone dataes respectively. We plot the obtained train loss and test loss to observe the experimental process during epochs in **Fig.6**. The loss curves of the training process decrease steadily, while the test loss curves fluctuate slightly but tend to decline steadily.

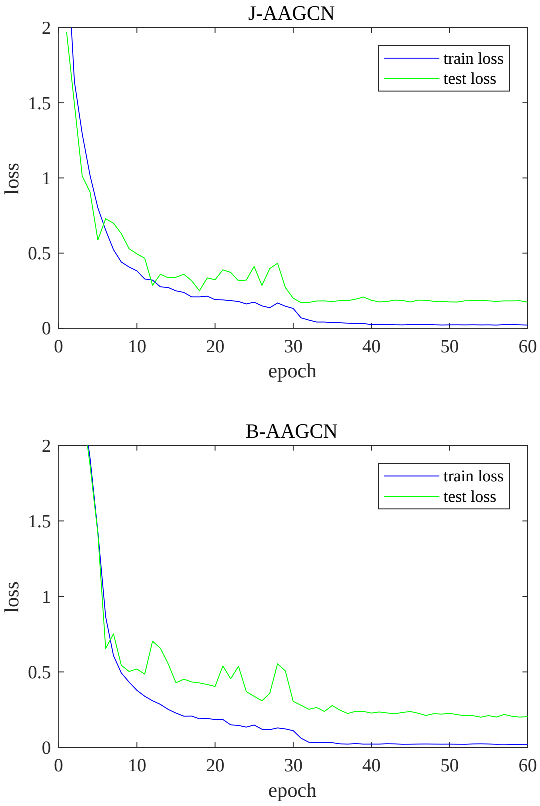

**Fig. 6.** Loss curves of our method.

For the filtered dataset, we respectively complete comparison experiments with these based methods, such as ST-GCN and 2s-AGCN. We compare the accuracy of our method with the above methods. It can be seen that our method has an improved effect on violent action recognition. The accuracy is shown in **Table 1**.

**Table 1.** The accuracy of several compared graph convolution methods on filtered NTU RGB+D dataset.

| Methods | Accuracy(%) |
|---|---|
| ST-GCN [15] | 81.54 |
| J-AGCN | 93.75 |
| B-AGCN | 92.77 |
| 2s-AGCN [20] | 94.76 |
| **Our method** | **Accuracy(%)** |
| J-AAGCN | 94.37 |
| B-AAGCN | 93.64 |
| 2s-AAGCN | 96.07 |

## 5. CONCLUSION

In this paper, we adopt a method of action recognition based on skeletal information extracted from campus surveillance video to detect violent actions. We use the graph convolutional network to process the skeleton data and get the classification score, so as to determine whether the video contained violent actions. We add the attention module into the two stream adaptive graph convolutional network. It is involved in the computation of adjacency matrices, which can improve the network's ability to extract spatial features. The comparison experiments based on the filter dataset show that our method is more accurate, it can be effective to determine some violent actions. Accordingly, if given other datasets, our method also can be used to detect other aggressive actions except punching, kicking and so on. Our method is based on human key point information, which is independent of the event scenario. Therefore, after training in one scenario, this method can be extended to numerous other surveillance devices in the campus.

In subsequent work, we can try to combine traditional image information with skeletal action recognition. It can be further optimized by incorporating other information, such as facial features. Our method is focus on the skeletal information as the spatiotemporal feature map, we can try to combine with other spatiotemporal information about objects in the surveillance video, such as stones, sticks and knives. This can help us conprehensively identify violent actions, so as to improve accuracy of the violent action recognition.

**Acknowledgements**

This work was supported by the Beijing Municipal Natural Science Foundation under Grant No. 3192028.

**References:**

[1] Nguyen D T, Li W, Ogunbona P O. Human detection from images and videos: A survey[J]. Pattern Recognition, 2016, 51: 148-175.

[2] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]. European Conference on Computer Vision. Springer, Cham, 2016: 20-36.

[3] Zhang H B, Lei Q, Zhong B N, et al. A survey on human pose estimation[J]. Intelligent Automation and Soft Computing, 2016, 22(3): 483-489.

[4] Liu J, Shahroudy A, Xu D, et al. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition[C]. European Conference on Computer Vision, 2016: 816-833.

[5] Pazhoumanddar H, Lam C P, Masek M, et al. Joint movement similarities for robust 3D action recognition using skeletal data[J]. Journal of Visual Communication and Image Representation, 2015: 10-21.

[6] Keceli A S, Can A B. Recognition of basic human actions using depth information[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2014, 28(02).

[7] Du Y, Wang W, Wang L, et al. Hierarchical recurrent neural network for skeleton based action recognition[C]. Computer Vision and Pattern Recognition, 2015: 1110-1118.

[8] Song S, Lan C, Xing J, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]. National Conference on Artificial Intelligence, 2017: 4263-4270.

[9] Zhang P, Lan C, Xing J, et al. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data[C]. International Conference on Computer Vision, 2017: 2136-2145.

[10] Liu J, Shahroudy A, Xu D, et al. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 3007-3021.

[11] Kim T S, Reiter A. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks[C]. Computer Vision and Pattern Recognition, 2017: 1623-1631.

[12] Liu M, Liu H, Chen C, et al. Enhanced skeleton visualization for view invariant human action recognition[J]. Pattern Recognition, 2017, 68(68): 346-362.

[13] Li C, Zhong Q, Xie D, et al. Skeleton-based action recognition with convolutional neural networks[C]. International Conference on Multimedia and Expo, 2017: 597-600.

[14] Li C, Zhong Q, Xie D, et al. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation[C]. International Joint Conference on Artificial Intelligence, 2018: 786-792.

[15] Yan S, Xiong Y, Lin D, et al. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[C]. National Conference on Artificial Intelligence, 2018: 7444-7452.

[16] Li C, Cui Z, Zheng W, et al. Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition[C]. National Conference on Artificial Intelligence, 2018: 3482-3489.

[17] Wei S, Ramakrishna V, Kanade T, et al. Convolutional Pose Machines[C]. Computer Vision and Pattern Recognition, 2016: 4724-4732.

[18] Chen Y, Wang Z, Peng Y, et al. Cascaded Pyramid Network for Multi-person Pose Estimation[C]. Computer Vision and Pattern Recognition, 2018: 7103-7112.

[19] Cao Z, Simon T, Wei S, et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields[C]. Computer Vision and Pattern Recognition, 2017: 1302-1310.

[20] Shi L, Zhang Y, Cheng J, et al. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition[C]. Computer Vision and Pattern Recognition, 2019: 12026-12035.

[21] Wang X, Girshick R, Gupta A, et al. Non-local Neural Networks[C]. Computer Vision and Pattern Recognition, 2018: 7794-7803.

[22] Shahroudy A, Liu J, Ng T, et al. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis[C]. Computer Vision and Pattern Recognition, 2016: 1010-1019.