Behavior Recognition Algorithm Based on Key Points of Human Bones

Yong Li
Chongqing University of Posts and Telecommunications
Key Laboratory of Industrial Internet of Things and Network Control, Ministry of Education
Chongqing, China
liyong@cqupt.edu.cn

Ziqiang Zhao
Chongqing University of Posts and Telecommunications
Key Laboratory of Industrial Internet of Things and Network Control, Ministry of Education
Chongqing, China
869186979@qq.com

2021 China Automation Congress (CAC) | 978-1-6654-2647-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAC53003.2021.9728615

***Abstract*—Human behavior recognition method based on video image easily influenced by background factors with lead to recognition accuracy is not high, and in view of the traditional Shuangliu network needs computing optical flow chart in advance and consume large amounts of time and need a lot of space to store an image of a light flow, an improved fusion skeleton information and image information of human behavior recognition algorithm. The experimental results show that the algorithm can effectively improve the accuracy of behavior recognition in video, and improve the identification ability of time-dependent behaviors and approximate behaviors.**

Keywords—Shuangliu network, Behavior recognition, Skeleton information, Image information

I. INTRODUCTION

In recent years, machine vision is developing rapidly, human behavior recognition, as a key part of video understanding, has always been one of the key research hotspots of vision researchers. It has high application value in the fields of human-computer interaction, virtual reality, intelligent human-computer interface and social video recommendation[1][2]. Because of the complex background, the appearent difference of objects and the similarity of different types of behaviors in real scenes, behavior recognition is still a challenging topic.

There are two main types of human behavior recognition [3]methods, traditional methods[4][5] and deep learning-based methods. The latter shows better performance, in which the dual-stream convolutional network method can effectively extract the apparent information and motion information from the video and achieve a good recognition effect in the behavior recognition task, but it is still difficult to effectively use the Spatio-temporal information in the video.

Human skeleton behavior recognition based on deep learning is a new method, which takes human skeleton in the video as input data and combines human skeleton information with other different algorithms to realize human behavior recognition. Using the key points of human bones to recognize human behavior has many advantages lies in that the model is obtained by learning the motion mode of human trunk. It pays more attention to the action itself, can eliminate the interference of image background information, and is not affected by illumination and other environments, and has strong robustness. However, focusing only on skeleton information brings disadvantages as well as advantages. Among all kinds of human behaviors, the action features of different behaviors often have similar places. When only skeleton information is used as the input feature, the computer may misjudge.

Based on the analysis of the above, this paper draw lessons from the Shuangliu method[6] is put forward in the original based on the human body skeleton on the convolution model of space and time, and one for the extraction of video frame image convolution branch of scene information and color information, this model can not only rely on the skeleton movement to identify the behavior, but also refer to the video frame image information, such as sports venues, the characters, clothing, etc., Identifying behaviors more accurately.

II. STRUCTURE OF SKELETON FLOW NETWORK

A. Methods an overview

The flow chart of the behavior recognition method based on human bone nodes is shown in Figure 1. This method firstly uses the improved OpenPose human 2D pose estimation model to extract the skeleton node coordinates of the human body in the video frame image, and then the obtained node coordinates are sent to the Inception network and the LSTM network simultaneously to extract the time and space information respectively. Finally, action classification results are obtained by a feature fusion algorithm.

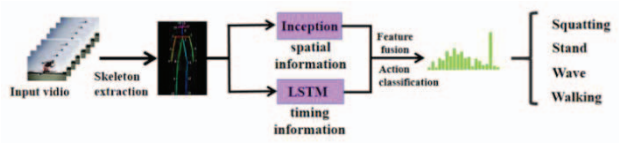

Fig. 1. Flow chart of behavior recognition model based on key points of human bones

B. Obtain target bone point

Bone key point detection module mainly detects important key points of the human body, such as key points, five sense nodes, etc. The OpenPose algorithm[7] uses a method called Part Affinity Fields (PAFS). Different from the previous human posture estimation algorithms, PAFS is a bottom-up posture estimation algorithm, and the previous posture estimation algorithms are top-down. The top-down algorithm is to detect people first and then key points, while openpose first detects the key points of human bones, then constructs the overall human skeleton, and finally carries out human posture recognition.

OpenPose has two output modes of human bone nodes. This paper only adopts the output format of 18 nodes, and the schematic diagram of the nodes is shown in Figure 2.

In fact, OpenPose can collect far more than 18 coordinates of key nodes, but this paper only uses 18 key locations related to posing recognition. The corresponding

relationship between node numbers from 0 to 17 is shown in Table 1.

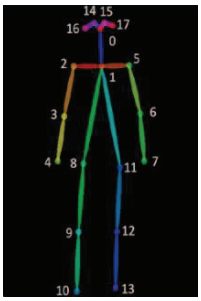

Fig. 2. Schematic diagram of key points of human skeleton

TABLE I. Key point labels correspond to tables

Serial number	Position	Serial number	Position
0	Nose	9	Right knee
1	Neck	10	Right ankle
2	Right shoulder	11	Left hip
3	Right elbow	12	Left knee
4	Right wrist	13	Left ankle
5	Left shoulder	14	Right eye
6	Left elbow	15	Left eye

The network structure of the OpenPose algorithm is shown in Figure 3. The first stage: The characteristic figure F is obtained from the input picture through the first 10 layers of vgg19 network; the second stage: The obtained feature map is used as the input of the 2-branch multi-stage convolutional neural network, in which the upper branch is used to predict a set of 2D part confidence maps of the position of body parts, while the lower branch is used to predict a set of 2D vector fields of partial affinity. Displays the Part Affinity Fields (PAF) between related nodes. The internal structures of L and S are shown in Fig.4.

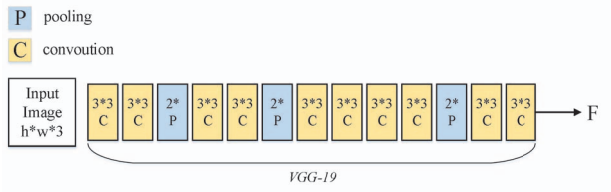

Fig. 3. VGG network structure diagram

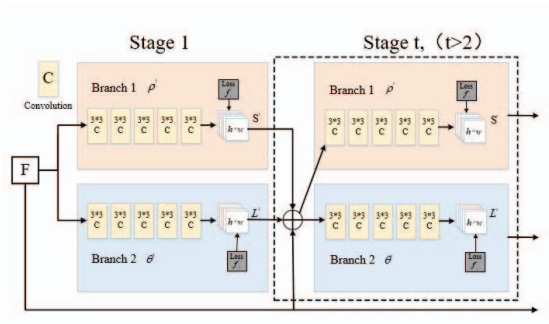

Fig. 4. OpenPose model internal structure diagram

For an input image with the size of $w \times h$, it first carries out feature extraction through a convolutional network to generate feature figure F, and then the feature figure F is input into the double-branch network. After a series of CNN processing, the 2D confidence graph S^1 and partial affinity L^1 of the node are obtained. The node confidence graph S and the body part affinity field L are respectively represented by the following formulae (1) and (2).

$$S^1 = \rho^1(F) \tag{1}$$

$$L^1 = \varphi^1(F) \tag{2}$$

Where ρ^1 and φ^1 represent the forward calculation of the first stage of the network. In the subsequent stage, the prediction results of the two branches of the previous stage and the original feature F are cascaded together as input. Its formulas are as (3) and (4).

$$S^t = \rho^t\left(S^{t-1}, L^{t-1}, F\right), \forall t \geq 2 \tag{3}$$

$$L^t = \varphi^t\left(S^{t-1}, L^{t-1}, F\right), \forall t \geq 2 \tag{4}$$

Through repeated iteration of a multi-stage convolutional neural network until the network converges. In the actual pose estimation of OpenPose, the image is finally obtained through the convolutional network to obtain the confidence graph of the node and the affinity field of the body part. In order to judge whether the two links in the confidence graph of the two links can be connected into a limb, the degree of alignment between the two links and the corresponding line segment in the affinity field of human body parts can be calculated. Specifically, d_{j1} and d_{j2} are defined as the coordinates of the two skeletal joint nodes detected by the confidence graph of the joint, and then they can be connected into the credibility of the limb E as:

$$E = \int_{u=0}^{u=1} L_c\left(p(u)\right) \frac{d_{j2} - d_{j1}}{\left\| d_{j2} - d_{j1} \right\|_2} du \tag{5}$$

Where represents the pixel points between continuous pixel points d_{j1} and d_{j2}, as shown in Equation (6) below.

$$p(u) = (1-u)d_{j1} + u d_{j2} \tag{6}$$

OpenPose feature extraction of the original model with VGG19 convolution neural network, the research shows that when the convolutional neural network after reaching a certain depth, not only fail to improve performance, it will cause the network convergence speed is slow, detection performance, and with the deepening of network layer number, and quantity will be more and more, computing speed drops. To address this shortcoming, Howard proposed MobileNet[8], a lightweight mobile terminal network with deep separable convolution as the core in 2017. Compared with the standard convolutional neural network, MobileNet has fewer parameters and lower computing costs.

The mobilenet model is based on deep separable convolution, which is a form of deconvolution. It decomposes standard convolution into deep convolution and 1×1 convolution. The structure of standard convolution and depth separable convolution is shown in the figure below. Assuming that the convolution kernel size is $D_k \times D_k \times M$, the parameters of standard convolution in Fig. 5 are as follows.

$$W_s = \left(D_k \times D_k \times M\right) \times N \tag{7}$$

While the convolution kernel size of deep convolution in Fig. 6 is $D_k \times D_k \times M$, with a total of M, and that of point convolution in Fig. 7 is $1\times1\times M$, with a total of N, the

parameters of deep convolution W_d and point convolution W_p are as follows.

$$W_d = (D_k \times D_k \times 1) \times M \tag{8}$$

$$W_p = (1 \times 1 \times M) \times N \tag{9}$$

The combination of depth direction convolution and point convolution is called depth direction separable convolution, so the parameter quantity W_D of depth direction separable convolution is the sum of depth direction convolution W_d and point convolution W_p.

$$W_D = W_d + W_p = (D_k \times D_k \times 1) \times M + (1 \times 1 \times M) \times N \tag{10}$$

Therefore, the ratio of the parameters of the depth separable convolution parameter quantity W_D to the standard convolution parameter quantity W_s is as follows:

$$\eta = \frac{W_D}{W_s} = \frac{1}{N} + \frac{1}{D_k^2} \tag{11}$$

The convolution core size of VGG19 is 3×3. Therefore, if we replace VGG19 with a mobilenet network with a deep separable convolution core, the number of parameters will be reduced by about 1/9.

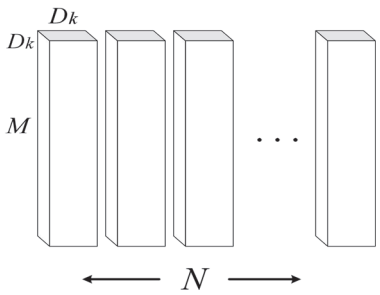

Fig. 5. Standard convolution

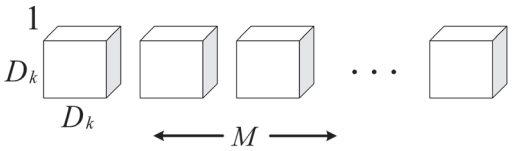

Fig. 6. Deep convolution

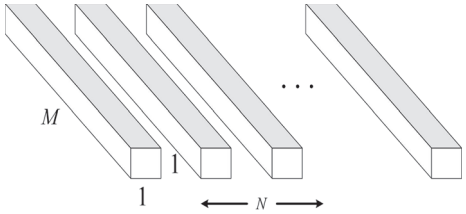

Fig. 7. Point convolution

C. Convolutional neural network

Because the video contains a lot of spatial and temporal information, spatial information exists in a single video frame, mainly including the scene information and the appearance information of the moving target, while temporal information exists between consecutive video frames, mainly including the motion information of the target. A convolutional neural network can well extract the spatial features of video frames. In this paper, the Google perception net is used to extract the spatial information of skeleton data set. The structure of the perception network is composed of one module, and the network structure is shown in Figure 8 below.

The concept module stacks the convolution kernels of 1×1, 3×3, 5×5 and their pooling together. On the one hand, it increases the width of the network, and on the other hand, it increases the adaptability of the network to scale. The convolution of the input by the first branch is a very excellent structure. At the same time, the dimension of the output channel can be increased and reduced. It can be seen that 1×1 convolution is used in all four branches of the inception module for low-cost (much less computation than 3×3) cross-channel feature transformation. The second, third and fourth branches are the same.

Compared with other convolution networks, the perception net achieves good classification performance while controlling the amount of computation and parameters; The last fully connected layer is removed and the global average pooling layer is used; Use the perception module to improve the utilization of parameters; Different convolution kernels are used to increase the diversity, so in this paper, the concept network is used to extract the spatial information of skeleton dataset.

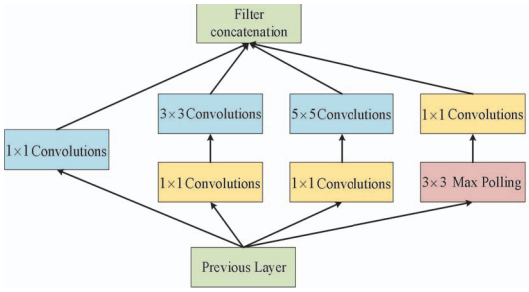

Fig. 8. Introduction module network structure diagram

D. Long term short term memory network

Long short-term memory (LSTM)[9] is a special recurrent neural network, which can effectively solve the problem of long-term data dependence[10]. Skeleton data can be regarded as the time-series information of human behavior. LSTM network is used to mine the context dependence of each limb behavior in a time domain and extract significant motion features, which can effectively improve the performance of human skeleton behavior recognition. The internal structure of the LSTM cell is shown in Figure 9.

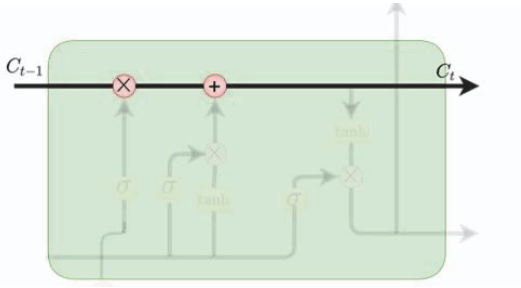

Fig. 9. Internal structure of LSTM cell

The research found that if only the above horizontal line is not able to add or delete information. To solve this

problem, LSTM is implemented through a structure called gates. In the gate structure, the sigmoid neural layer and point by point multiplication method are introduced to orderly let the information pass through the horizontal line. As shown in Figure 10, it is a single information node diagram of the door structure.

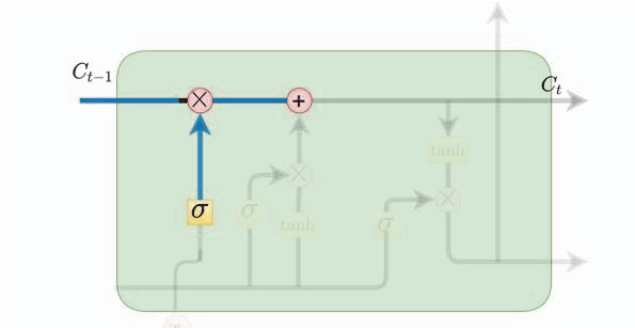

Fig. 10. Information node diagram

Any element output by the sigmoid layer is a real number between 0 and 1, which represents the weight of the corresponding information passing. For example, 0 means "no information is allowed to pass" and 1 means "all information is allowed to pass". Compared with RNN algorithm, three gates with the same structure are added to the network structure of LSTM algorithm. In LSTM algorithm, the three gates are conducive to the protection and control of input and output information.

1) Forgetting gate.

Based on the above, some information will be lost when passing through the horizontal line, and this process needs to introduce a forgetting gate to discard this part of information. The forgetting gate will input the last time output value h_{t-1} and the current network input value x_t in LSTM at the same time, and the real number between 0 and 1 will be output after passing the forgetting gate. The output real number is assigned to each number in neuron state C_{t-1}. If the neuron state is 1, it means "completely reserved"; 0 stands for "discard all". The network structure of forgetting gate is shown in Figure 11.

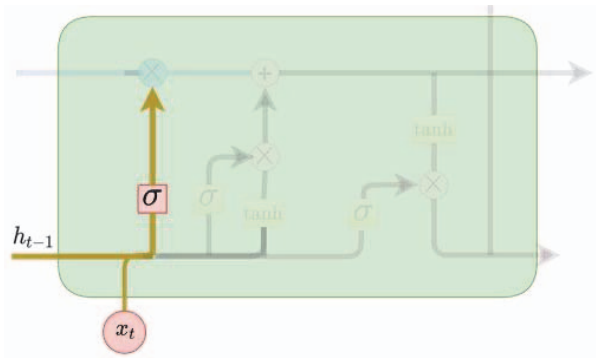

Fig. 11. Forgetting gate network structure diagram

2) Input gate.

In addition to the information that the forgetting gate needs to discard, some new information needs to be added to the neuron state. In order to add new information to neuron state, LSTM algorithm introduces the concept of input gate. On the one hand, the sigmoid layer of the input gate can determine which information needs to be updated; On the other hand, each tanh in the network structure can generate a vector, which is used to alternative information that needs to be updated. Fig. 12 shows the network structure of input gate.

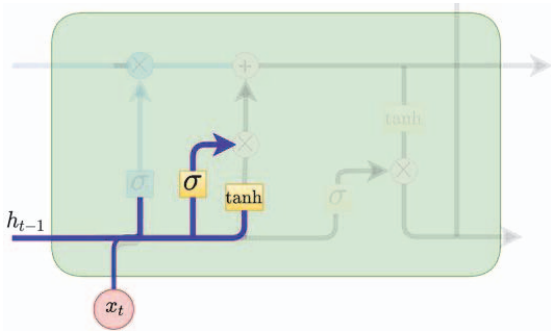

Fig. 12. Network structure diagram of input gate

According to the network structure of LSTM input gate, LSTM needs to update the time of old neuron state, and update C_{t-1} to C_t. In order to update the state of the neuron, it is necessary to multiply the output value h_{t-1} of LSTM at the last time by the current network input value x_t, and discard the information that really needs to be discarded by using the forgetting gate. On this basis, $i_t * \widetilde{C}_t$ is added as a new candidate value, which changes according to the degree of each state update.

1) Output gate.

In order to improve the network structure of LSTM algorithm, we need to introduce the concept of output gate. The output in the output gate will be based on the updated neuron state, which is the updated neuron state. First, the sigmoid layer operation is used to determine which part of the neuron will be output. Then, Neurons passing through the tanh layer will get a value between - 1 and 1. After multiplication, only the output part that has been determined before is output. Figure 13 shows the network structure of the output gate.

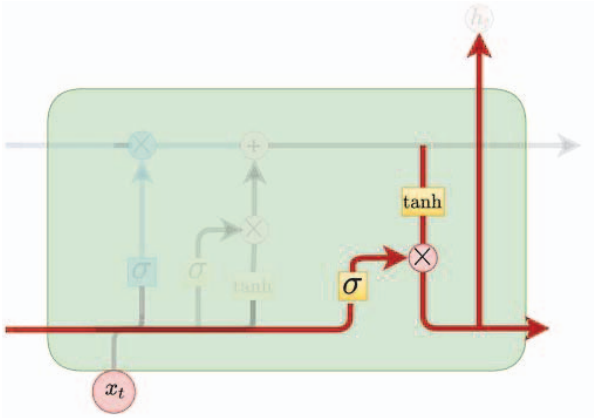

Fig. 13. Schematic diagram of output gate network structure

III. Image Stream Network Structure

A. Time segment network

Behavior recognition based on human skeleton information only focuses on the skeleton movement. The operation of extracting skeleton information from video skeleton is relatively simple and real-time. The amount of data input by skeleton data is far less than the space occupied by optical flow. However, focusing only on skeleton

information not only reflects its advantages, but also brings its disadvantages. Among various human behaviors, similar actions often exist between different behaviors, Such as sweeping the floor and Tai Chi. They all have a lot of movements such as moving and jumping. When only skeleton information is used as the input feature, the computer may misjudge. Therefore, on this basis, an image convolution branch for extracting video frame scene information and color information is added, In this way, the model can not only recognize the behavior with the skeleton information, but also recognize the image information of video frames, such as scene information, character clothing, etc., so as to improve the accuracy of recognition.

In view of the poor modeling ability of traditional two stream method for long-time video data, time segment network can effectively solve the problem of poor modeling ability of traditional two stream model for long-time video data. The schematic diagram of TSN network is shown in Figure 14.

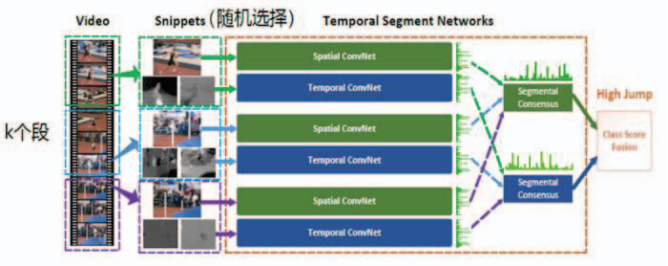

Fig. 14. TSN network structure diagram

As can be seen from the above figure, the original TSN network also adopts the idea of double flow method, and the time flow also needs to calculate the optical flow diagram of video frames, which is a large amount of calculation and a waste of space. Research shows that deeper network structure can improve the performance of object recognition. However, the original dual stream network model uses a relatively shallow network structure, so this paper uses a deeper resnet-101 convolutional neural network to extract the feature information of each video frame, The whole input video is segmented by sparse sampling strategy, and the video frames are randomly sampled from each segment as the input of image classification network. Each video frame is classified, and the score of each segment is obtained. The classification results of each segment are preliminarily fused by segment consensus function, and the video level classification results of the original video are obtained. The improved TSN network structure is shown in Figure 15.

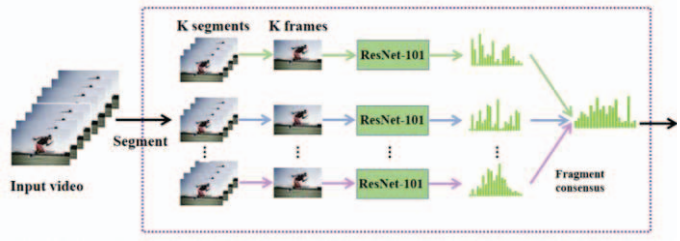

Fig. 15. Improved TSN network structure

Finally, image stream and skeleton stream are fused by concat fusion to get the final classification result of the whole model.

IV. EXPERIMENTAL RESULT

A. Experimental environment

The simulation environment is shown in Table 2.

TABLE II. Introduction of experimental environment

	Name	Introduce
Hardware environment	Process	Intel Core i5-10600KF@4.80GHz
	Graphics card	NVIDIA GeForce GTX 1080
	Memory	16.0 GB
Software environment	Operating system	Windows 10 Enterprise Edition
	Development language	Python
	Development environment	Python3.6+CUDA9.0+TensorFlow-GPU1.12.0

B. dataset

For behavior recognition databases, the most common ones are MS coco, Weizmann dataset , ucf11, UCF sports action dataset, ucf50 and ucf101. This paper uses coco dataset, which is a large and rich object detection, segmentation and subtitle data set. There are 80 categories, more than 330000 images, of which 200000 are labeled. The number of individuals in the whole data set is more than 1.5 million. Since this paper only recognizes human posture, 5000 pictures of human posture are selected and divided into four categories: squatting, standing, waving and walking. 4500 of them are used as training set and 500 as test set.

C. Network training and parameter description

In this paper, the above data set is used as the training data set of human bone joint points, and a model suitable for learners' behavior recognition is obtained through training. The relevant experimental parameters are shown in Table 3.

TABLE III. Description of relevant experimental parameters

Parameter name	Parameter value
Input size	656×368
Epoch	10
Batch_size	32
Learning_rate	0.0001

D. Comparison of experimental results

In this paper, we improve the feature extraction network in OpenPose multi person pose estimation algorithm. By replacing vgg19 feature extraction network with mobile thin, we add another branch image stream network structure to extract the scene information in the video based on skeleton stream network structure to extract human pose in the video, and fuse the recognition results of the two branch networks to get the final recognition result. The experimental results are shown in Table 4, the curves of train acc, train loss, val acc and val loss are shown in Figure 16 (before left improvement and after right improvement), and the confusion matrix is shown in Figure 17 (before left improvement and after right improvement).

It can be seen from table 4 that the FPS of the network structure before improvement can only reach 9 ~ 10, while the FPS after improvement can reach more than 25. The recognition accuracy of the network structure before improvement is 0.94, and the recognition accuracy of the network structure after improvement is 0.97. The experimental results show that the detection speed of the improved network is improved by about 60% while the

accuracy is improved. Therefore, the improved network is more suitable for the actual scene.

TABLE IV. Comparison before and after improvement

Object detection algorithm	Accuracy	FPS
Original network	0.94	9
Improved network	0.97	26

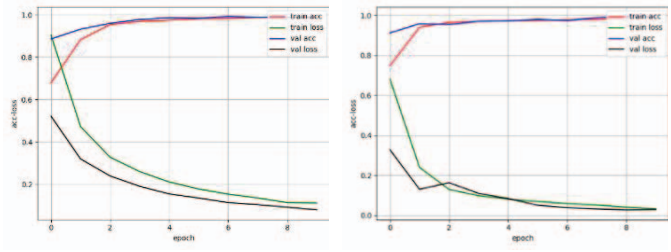

Fig. 16. Train acc,train loss,val acc,val loss diagram

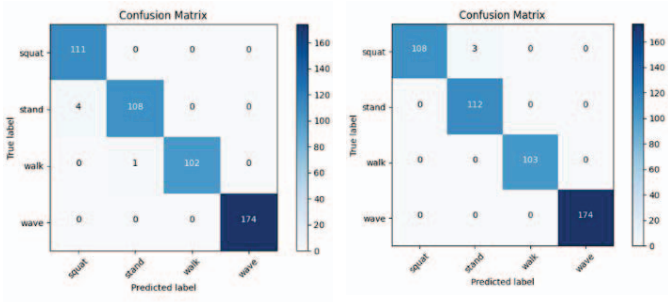

Fig. 17. Confusion matrix

V. CONCLUSION

In this paper, the human skeleton key point behavior recognition algorithm is studied. A human behavior recognition algorithm based on improved two stream spatiotemporal network is proposed. In one branch, the key points of skeleton are used to effectively extract the temporal and spatial features of human body in video. In the other hand, the convolution network of two branch image stream is introduced to extract the spatial information of scene in video, which further improves the recognition accuracy, The experimental results show that: the recognition accuracy of this algorithm reaches 97%, which is higher than the existing algorithms, and it is 60% higher than the previous network structure FPS. The next research will further expand the data set, especially in complex scenes, while ensuring the accuracy, and further optimize the model to improve the detection speed.

ACKNOWLEDGMENT

This work was supported by Technology Innovation and Application Development Project of Chongqing (cstc2019jscx-fxydX0060).

REFERENCES

[1] B. Fu, N. Damer, F. Kirchbuchner and A. Kuijper, "Sensing Technology for Human Activity Recognition: A Comprehensive Survey," in IEEE Access, vol. 8, pp. 83791-83820, 2020.

[2] A. V. Vesa et al., "Human Activity Recognition using Smartphone Sensors and Beacon-based Indoor Localization for Ambient Assisted Living Systems,"2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 2020, pp. 205-212.

[3] L. Shao, D. Wu and X. Li, "Learning Deep and Wide: A Spectral Method for Learning Deep Networks," in IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 12, pp. 2303-2308, Dec. 2014.

[4] L. Xia, C. Chen and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 2012, pp. 20-27.

[5] R. Vemulapalli, F. Arrate and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014 pp. 588-595.

[6] W. Ye, J. Cheng, F. Yang and Y. Xu, "Two-Stream Convolutional Network for Improving Activity Recognition Using Convolutional Long Short-Term Memory Networks," in IEEE Access, vol. 7, pp. 67772-67780, 2019.

[7] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021.

[8] Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV. Searching for mobilenetv3. InProceedings of the IEEE International Conference on Computer Vision 2019 (pp. 1314-1324).

[9] B. Praveen Kumar and K. Hariharan, "Multivariate Time Series Traffic Forecast with Long Short Term Memory based Deep Learning Model," 2020 International Conference on Power, Instrumentation, Control and Computing (PICC), Thrissur, India, 2020, pp. 1-5.

[10] Y. Yu, X. Si, C. Hu and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," in Neural Computation, vol. 31, no. 7, pp. 1235-1270, July 2019.