ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 146

2nd International Conference on Artificial Intelligence: Technologies and Applications (ICAITA 2018)

Research of Action Recognition Methods Based on RGB+D Videos

Zhongyin Huang* and Wei Chen

School of Electronic Information Engineering, Beihang University, Beijing, China

*Corresponding author

***Abstract*—In order to solve the problem on making full use of RGB+D dataset that includes RGB data, 3D skeletal data, depth map sequences and infrared videos, this paper proposes an action recognition method of RGB+D videos that merges a multi-layer recurrent neural network and two-stream convolutional networks, combining RGB information and joints information together. Simulation results show that the multi-layer recurrent network proposed in this paper has better performance than other recurrent networks when dealing with the skeletal data. Moreover, by combining it with the spatial network or temporal network through nonlinear weighted score fusion, the recognition accuracy is further improved. The cross-view action recognition accuracy is improved to be 0.79%, 5.6%, 20.62% and 23.65% higher than the original method, respectively by using the multi-layer network alone, combining the multi-layer network and spatial network, combining the multi-layer network and temporal network, and combining three networks together.**

Keywords—action recognition; RGB+D; TSN; RNN

I. INTRODUCTION

In recent years, methods of action recognition based on deep learning have been received much interesting. One of them is the two-stream convolutional networks that outperform many previous methods. Simonyan et al. [1] proposed the two-stream convolutional networks by merging the spatial and temporal information based on the convolutional networks. In [2], the recognition performance of merging spatial networks and temporal networks in different levels was discussed and the two-stream convolutional networks were further improved. Wang et al. [3] combined iDT(improved dense trajectories) with the two-stream networks. They used TDDs(trajectory-pooled deep-convolutional descriptors) to describe features and encoded it into a high dimensional representation by Fisher vector. The two-stream convolutional networks are recognition model mainly aimed at RGB videos. However, the data collected by the Kinect camera not only include RGB data, but also include 3D skeletal data, depth map sequences and infrared videos. Chen et al. [4] used a depth camera and an inertial sensor to built a real time action recognition system by a decision-level fusion. Wang et al. [5] proposed a algorithm to mine a set of key-pose-motifs to recognize actions from skeletal data by matching a sequence to the motifs of each class and selecting the class that maximizes the matching score. Shahroudy et al. [6] built the NTU RGB+D dataset and proposed the P-LSTM(part-aware long short-term memory) network for action recognition from skeletal data.

In most of previous researches, either depth map or skeletal data was used alone for action recognition in RGB+D videos. Shahroudy et al. [7] proposed a shared-specific feature factorization network to separate input multimodal signals into a hierarchy of components. This network achieved much higher accuracy in action recognition of RGB+D videos, but the result is not ideal enough because of the poor performance of the RGB based features for the cross-view task. Therefore, this paper focuses on how to make full use of the data collected by Kinect camera to recognize human actions in cross-view RGB+D videos. The RGB data and skeletal data are chosen out of four types of data provided. For RGB data, the TSN [8] (temporal segment networks) model is applied and the inputs of the model are RGB frames and optical flow stacks extracted from RGB videos. As for skeletal data, a multi-layer recurrent network is proposed. Finally, a new TSN model(TSN2) is built by merging the results of different input modalities through weighted prediction score fusion.

II. TSN2 MODEL

In this paper, the proposed TSN2 model is constructed by two parts. One is the two-stream convolutional network [8], the structure of which is BN-Inception [9] that initialized by model pre-trained on Kinetics dataset. The other is the multi-layer recurrent network for skeletal data processing. Both of them adopt the temporal segment method to deal with input data and use score fusion of output results for classification and recognition.

A. Two-stream Convolutional Neural Networks

The basic structure of TSN model proposed in [8] is two-stream convolutional neural networks. Two-stream networks [1] includes two convolutional networks: spatial network and temporal network, combining spatial and temporal information. The internal structure of the two networks is basically the same BN-Inception network. The main different is the input layer. During training, two networks are trained separately by RGB or optical flow images from training set. During testing, RGB or optical flow images of testing set are put into the already trained spatial network or temporal network respectively.

B. Multi-layer Recurrent Neural Network

To deal with the skeletal data, another stream is added besides the two streams in TSN model proposed in [8]. Shahroudy et al. [6] proposed a P-LSTM (part-aware long short-term memory) network to recognize actions. Drawing on the experience of this method, a multi-layer recurrent network that consists of one LSTM [10][11][12] layer followed by two GRU [13] layers is proposed in this paper and proved to

improve the recognition accuracy. The configuration of multi-layer recurrent network is shown as Figure 1.

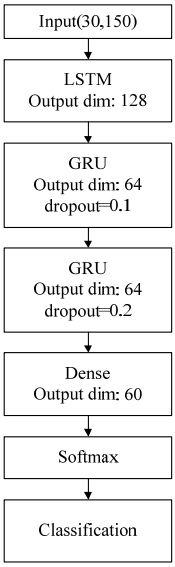

FIGURE I. CONFIGURATION OF MULTI-LAYER RECURRENT NEURAL NETWORK

III. TRAINING TSN2 MODEL

A. RGB+D Dataset

The research in this paper used the NTU RGB+D Action Recognition Dataset [6] made available by the ROSE Lab at the Nanyang Technological University, Singapore. The dataset was collected by Microsoft Kinect v2 sensor, totally 56880 action samples including RGB videos, depth map sequences, 3D skeletal data, and infrared videos for each sample. The configuration of body joints is shown in Figure 2, and an example frame of RGB videos is illustrated in Figure 3.

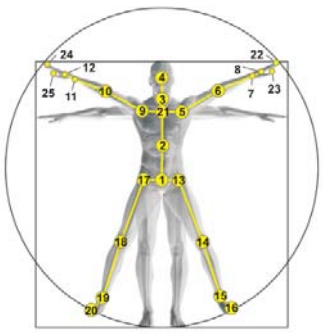

FIGURE II. CONFIGURATION OF BODY JOINTS IN 3D SKELETAL DATA.

FIGURE III. AN EXAMPLE FRAME OF RGB VIDEOS

B. Input Data Processing

In this paper, the research mainly aimed at cross-view action recognition. Therefore, the videos captured by camera 2 and 3 are assigned as training set, including 37920 videos. The videos captured by camera 1 are assigned as testing set with 18960 videos.

For the input of the two-stream convolutional networks, we follow the method same as [8]. As for the input of multi-layer recurrent network, since there are action samples performed by a single person or several people, the skeletal data saved by cameras also include one skeleton or several skeletons for a sample, even noisy skeletons that are actually tables or chairs. We use similar method as [6] to extract useful data from the provided skeletal data. The input of the network is set to two vectors. If the camera only detected one skeleton, the data are saved in the first vector and the second vector is set to zero. If two or more skeletons were detected, we first decide whether the maximal range of each skeleton on y axis is reasonable, and filter the noisy skeletons by setting a threshold. Secondly, the sum of variance of each joint's xyz coordinates in the rest skeletons is calculated. The skeleton with the largest variance is seen as the main skeleton and saved in the first vector, while the skeleton with second large variance is saved in the second vector. Since one or more skeleton may be detected in any frame of the video, the xyz coordinates will be set to zero in the frames that a skeleton is not detected by the camera. Finally, the two vectors are integrated into one vector. For example, the vector of frame 0 is supposed to be in the form of:

$$[f_0s_0x_0, f_0s_0y_0, f_0s_0z_0, f_0s_1x_0, f_0s_1y_0, f_0s_1z_0, ...,$$
$$f_0s_0x_{24}, f_0s_0y_{24}, f_0s_0z_{24}, f_0s_1x_{24}, f_0s_1y_{24}, f_0s_1z_{24}]$$

$f_is_jx_k$ represents the x coordinates of joint k in skeleton j, frame i.

We calculate the range of each skeleton on y axis by calculating the absolute value of the subtraction between the y coordinates of head(No.4 joint) and right foot(No.20 joint). The threshold is set according to several videos that only a single skeleton is detected.

In the experiment, K=3. All RGB, optical flow and skeletal data are adopted similar temporal segment method of processing the final inputs, dividing the data into 3 segments and randomly selecting one frame, five continuous frames and ten continuous frames respectively.

A. Training Parameters of the Model

When training spatial network and temporal network, the changes based on [8] are mainly the learning rate, iteration size, batch size and clip gradient.

When training multi-layer recurrent network, the optimization algorithm is Adam (Adaptive Moment Estimation) [14]. In this paper, the initial learning rate is 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Besides, we set a callback function to check the validation accuracy. When the validation accuracy is not improved in last three continuous iterations, the learning rate is reduced to its half and stops declining until 0.00001. The batch size is 32.

B. *Result of Training the Model*

We train the model with one GTX1080Ti GPU. Time of training the spatial network is about two days, and about a week on training temporal network, about two hours on training multi-layer recurrent network. The performances of three trained networks are shown in Table 1.

TABLE I. PERFORMANCE OF THREE NETWORKS IN TSN MODEL

Name of Networks	Performance	
	loss	***accuracy***
spatial network	0.436	0.762
Temporal network	0.520	0.836
Multi-layer recurrent network	0.744	0.767

IV. TESTING TSN2 MODEL

A. *Testing Result of Two-stream Network*

According to the method in [8], 25 RGB frames or 25 optical flow stacks (5 continuous frames in a stack) are selected in a video and put into spatial network or temporal network for testing. Then the average of 25 groups of output scores of the fully-connected layer are calculated. Finally the softmax scores are calculated for classification. The average accuracy of the two networks is 75.90% and 90.86% respectively. The recognition accuracy of spatial network is comparatively low in the following action classes. The 11th: reading, 0.491; the 19th: take off glasses, 0.427; the 29th: playing with phone/tablet, 0.443; the 34th: rub two hands together, 0.215; the 38th: salute, 0.478; the 39th: put the palms together, 0.304. The highest recognition accuracy is in the 59th: walking towards each other, 1.00. The recognition accuracy of temporal network is obviously higher than that of spatial network. The lowest action class is the 12th: writing, only 0.373; the highest is the 55th: hugging other person, the 59th: walking towards each other, and the 60th: walking apart from each other, 1.00.

B. *Testing Result of Multi-layer Recurrent Network*

The method of testing the multi-layer recurrent network is different from the two networks above. But similar to the training process, we divide each video into 3 segments, and randomly select 10 continuous frames, totally 30 frames, as the testing frames of this video. The average accuracy is 71.06%.

Apparently, the recognition accuracy is relatively lower than temporal network and spatial network. The action class with the lowest accuracy is the 12th: writing, only 0.267, which is similar to temporal network. The second lowest is the 11th: reading, 0.394. Besides, the 10th: clapping, 0.491; the 17th: take off a shoe, 0.484; the 29th: playing with phone/tablet, 0.453; the 34th: rub two hands together, 0.472; the 37th: wipe face, 0.440; the 47th: touch neck, 0.459, also have relatively lower accuracy. Actions with low recognition accuracy in three networks are relatively slight motions. Moreover, the action classes with higher accuracy are also similar to those two networks above: the interactive action pushing other person, 0.908; hugging other person, 0.971; walking towards each other, 0.930; walking apart from each other, 0.952. Furthermore, the 8th: sitting down, 0.917; the 9th: standing up, 0.968; the 22nd: cheer up, 0.917; the 43rd: falling, 0.978, are also easier to recognize because of relatively large range of movements and single movement direction. However, the multi-layer recurrent network has higher accuracy in some action classes although it performs worse in whole. For example, as for the accuracy of recognizing sitting down and standing up, multi-layer recurrent network is higher than spatial network.

A. *Merging the Testing Results of Three Networks*

Network fusion at two different positions: before softmax and after softmax; with two different algorithms: linear weighted and nonlinear weighted fusion, are studied in this paper.

Fusion before softmax is the method used in [8]. Combining three networks at this step is to multiply the output of each network with a weight and add them up to get a merged vector. Fusion after softmax is multiplying the obtained vector after softmax with a weight, and then adding the weighted result of three networks together.

As for the linear weighted fusion, it is selecting a scalar as the weight. While the nonlinear fusion is selecting a vector as the weight. The vector is called accuracy vector, constructed by testing a part of data and listing the accuracy of each class in the order of the action class labels. We multiply the output of each network before softmax or after softmax with the vector element wisely. The testing results of networks merged in different conditions are shown in Table 2. acc in the tables stands for accuracy vector.

TABLE II. TESTING RESULT OF NETWORKS MERGED IN DIFFERENT CONDITIONS

Fusion Position	Algorithm	Weight of Spatial Network	Weight of Temporal Network	Weight of Multi-layer recurrent Network	Fusion Result
Before Softmax	Linear Weighted	1	4	1	92.48%
		1	4	0	93.93%
		4	0	1	73.93%
		0	4	1	88.82%
	Nonlinear Weighted	1	3.7	acc	93.92%
		4	0	acc	75.87%
		0	4	acc	90.89%
After Softmax	Linear Weighted	1	4	1	92.28%
		1	4	0	92.54%
		4	0	1	75.65%
		0	4	1	90.25%
	Nonlinear Weighted	1	2.73	acc	92.81%
		1	0	acc	75.82%
		0	1	acc	90.87%

The accuracy of spatial and temporal network fusion before softmax with linear weight is the highest. The second highest is the nonlinear weighted fusion of three networks before softmax. The participation of multi-layer recurrent network makes the accuracy decline slightly. But if we only combine the multi-layer recurrent network with either spatial network or temporal network, the accuracy of merged network will increase, higher than the used networks before fusion. Therefore, in the condition that requires more about real-time capability, we can choose to merge the spatial network and multi-layer recurrent network that have higher operation speed. While in the

condition that requires more about accuracy and cares less about time consumption, we can choose the fusion of the two-stream networks. Besides, when the weight of multi-layer recurrent network is a scalar, its participation will reduce the accuracy of the original network. Thus we proposed the nonlinear weight according to its accuracy on different action classes and weight each dimension of the score vector before fusion separately. Experiments show that this method performs better than linear weighting but is not suitable for spatial or temporal network. The testing accuracy comparison between the method in this paper and in [6] is listed in Table 3. It can be seen that the multi-layer recurrent network(MLR) proposed in this paper outperforms P-LSTM in cross-view action recognition, and combining it with spatial network(SN) or temporal network(TN) achieves higher accuracy.

TABLE III. ACCURACY OF DIFFERENT NETWORKS

Name of Network	Cross-view Action Recognition
2 Layer RNN [6]	64.09%
2 Layer LSTM [6]	67.29%
2 Layer P-LSTM [6]	70.27%
MLR	71.06%
MLR+SN	75.87%
MLR+TN	90.89%
MLR+SN+TN	93.92%

V. CONCLUSIONS

In this paper, a new TSN model is proposed for action recognition in RGB+D videos. It combines two types of data: RGB videos and 3D skeletal data by merging multi-layer recurrent network and two-stream convolutional networks. The original TSN model is adopted for the RGB videos. RGB images and optical flow images are extracted as the inputs of two-stream convolutional networks. As for the 3D skeletal data, a multi-layer recurrent network is proposed in this paper, which learns the features in skeletal data for action recognition. Research on the fusion method of spatial network, temporal network and multi-layer recurrent network is conducted. Experiments show that merging three networks before softmax with nonlinear weights can better improve the recognition accuracy; the accuracies of multi-layer recurrent network alone or combined with other two networks are all higher than that in reference. However, the improvement made by participation of multi-layer recurrent network is not evident enough since its own accuracy is lower than the spatial network and temporal network. The further work is to improve the multi-layer recurrent network to make better use of skeletal data and further promote the action recognition accuracy.

REFERENCES

[1] K Simonyan, A Zisserman, “Two-stream convolutional networks for action recognition in videos,” Advances in Neural Information Processing Systems. 1(4):568-576, 2014.

[2] C Feichtenhofer, A Pinz, A Zisserman, “Convolutional Two-Stream Network Fusion for Video Action Recognition,” IEEE Conference on Computer Vision and Pattern Recognition. 2016:1933-1941.

[3] L Wang, Y Qiao, X Tang. “Action recognition with trajectory-pooled deep-convolutional descriptors,” IEEE Conference on Computer Vision and Pattern Recognition. Boston. 2015:4305-4314.

[4] C Chen, R Jafari, N Kehtarnavaz. “A real-time human action recognition system using depth and inertial sensor fusion,” IEEE Sensors Journal. 16(3):773-781, 2016.

[5] C Wang, Y Wang, A L Yuille. “Mining 3D key-pose-motifs for action recognition,” IEEE Conference on Computer Vision and Pattern Recognition. 2016:2639-2647.

[6] A Shahroudy, J Liu, T T Ng, G Wang. “NTU RGB+D: a large scale dataset for 3D human activity analysis,”IEEE Conference on Computer Vision and Pattern Recognition. 2016:1010-1019.

[7] A Shahroudy, T T Ng, Y Gong, G Wang. “Deep multimodal feature analysis for action recognition in RGB+D videos,” IEEE Transactions on Pattern Analysis & Machine Intelligence. PP(99):1-1, 2017.

[8] L Wang, Y J Xiong, Z Wang, Y Qiao, D H Lin, et al. “Temporal segment networks: towards good practices for deep action recognition,” Computer Vision ECCV 2016. 2016:20-36.

[9] S Ioffe, C Szegedy. “Batch normalization: accelerating deep network training by reducing internal covariate shift,” International Conference on International Conference on Machine Learning. 2015:448-456.

[10] H Sak, A Senior, F Beaufays. “Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition,” Computer Science. 2014:338-342.

[11] S Hochreiter, J Schmidhuber. “Long short-term memory,” Neural Computation. 9(8):1735-1780, 1997.

[12] H Sak, A Senior, F Beaufays. “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” Computer Science. 2014:338-342.

[13] J Chung, C Gulcehre, K Cho, Y Bengio. “Gated feedback recurrent neural networks,” Computer Science. 2015:2067-2075.

[14] D Kingma, J Ba. “Adam: a method for stochastic optimization.” International Conference on Learning Representations. 2015:1-13.