Journal of Korea Multimedia Society Vol. 24, No. 3, March 2021(pp. 423-430)
https://doi.org/10.9717/kmms.2020.24.3.423

Dual-Stream Fusion and Graph Convolutional Network for Skeleton-Based Action Recognition

Zeyuan Hu[†], Yiran Feng[††], Eung-Joo Lee[†††]

ABSTRACT

Aiming Graph convolutional networks (GCNs) have achieved outstanding performances on skeleton-based action recognition. However, several problems remain in existing GCN-based methods, and the problem of low recognition rate caused by single input data information has not been effectively solved. In this article, we propose a Dual-stream fusion method that combines video data and skeleton data. The two networks respectively identify skeleton data and video data and fuse the probabilities of the two outputs to achieve the effect of information fusion. Experiments on two large dataset, Kinetics and NTU-RGBC+D Human Action Dataset, illustrate that our proposed method achieves state-of-the-art. Compared with the traditional method, the recognition accuracy is improved better.

Key words: Skeleton-Based Action Recognition, Graph Convolutional Network, Skeleton Data, Video Data, Information Fusion

1. INTRODUCTION

Human motion recognition has a wide range of applications in video surveillance, healthcare, smart home, smart driving, and human-computer interaction [1,2]. At present, the task of action recognition mainly includes two directions, the method based on RGB video and the method based on skeleton data. However, RGB-based action recognition relies too much on the appearance information of the environment and objects. Problems such as background interference and sensitivity to lighting changes have always limited the development of RGB-based methods. By contrast, skeleton-based methods focus more on the action features themselves. In other words, the skeleton data contain the position information of human joints with strong robustness to scale changes and illumination changes, which also have invariance to the camera angle, human rotation and motion speed[3,4]. Deep learning automatically mines the information features in data through network fitting, which has become a new direction in the field of action recognition. For data sets, action recognition methods based on deep learning can be divided into recognition methods based on video data and recognition methods based on skeleton data.

The traditional action classification method for video data is based on the spatial information in the video frame and the information representation in the time domain. Simonyan et al.[1] proposed a dual-stream recognition method, which trains a convolutional neural network on video images and dense optical flow separately, then averages the input probabilities and outputs the final recognition rate. The dual-stream method can not only learn the color contour information of the video picture, but also learn the short-term time sequence in-

※ Corresponding Author : Eung-Joo Lee, Address: 428, Sinseon-ro, Nam-gu, Busan, Korea, TEL : +82-53-629-1143, FAX : +82-, E-mail : ejlee@tu.ac.kr
Receipt date : Jan. 19, 2021, Approval date : Mar. 15, 2021
[†] Dept. of Information Communication Engineering, Tongmyong University (E-mail : dlhzy410@126.com)
[††] Dept. of Information Communication Engineering, Tongmyong University
(E-mail : 345509139@qq.com)
[†††] Dept. of Information Communication Engineering, Tongmyong University

formation of the action, but it cannot handle the long-term time sequence information network of the action well.

Video data has good color information and contour information, which can be fully learned through the network model. Moreover, video data collection is convenient, and it is relatively easy to construct large-scale data sets. However, the video data collected by the camera only has the positional features of the two-dimensional plane. For video images, most of the current network models can only learn human color contour information, and the information richness is low, which is the reason for the low recognition accuracy of such methods.

The recognition method based on skeleton data achieves the purpose of classification by constructing a network model that can learn the spatial position information of the joint point coordinates and the time sequence change information of the joint points. However, the three-dimensional coordinates of the human joints are simply input into the network model as a whole, and the link relationship between the keys of the human body is not fully considered. However, the skeleton data has spatial three-dimensional information, which can directly reflect the spatial transformation of human motion. However, the contour information of the image is lacking. For actions with small differences, only bone data is used, which is often difficult for the network to distinguish.

Therefore, this paper rewards different network models for the two data respectively, and performs probabilistic fusion of the output results of the two models. This method combines the advantages of the two kinds of data to a certain extent, and effectively realizes the information fusion of the two kinds of data.

2. RELATED WORK

The operators in traditional CNNs, convolution and pooling, depend on the translation invariance of data. However, the graph data, as non-Euclidean data, do not satisfy translation invariance, and each node has a different local structure. Compared with the classic spatio-temporal convolutional neural network[5], GCN solves the problem of high computational complexity and high dimensionality of the generated features, which will take up a lot of memory. Bruna et al.[6] formulated the first graph convolution neural network in 2013, and they defined graph convolution in spectral space based on the convolution theorem. This method was later developed into the spectral method. Defferrard et al.[7] adopted recursive Chebyshev polynomials as filters, which is more efficient than the previous filters. To make graph convolution useful in the field of semi-supervised learning on graphs, Kipf and Welling[8] simplified Chebyshev networks. They proposed first-order graph convolutional neural networks[8], which contains aggregation functions that define node correlations. Thus, this method is seen as a bridge between the spectral and spatial methods. Chebyshev networks[7] and first-order graph convolutional neural networks[8] can be viewed as using Laplace matrices or its varieties as aggregation functions. The core idea of spatial graph convolution is aggregating node information using edge information to generate new node representations. Niepert et al.[9] use a greedy strategy to sort the nodes in a local neighborhood and then perform a spatial convolution filtering operation on the sorted nodes.

The convolutional neural network performs convolution calculations on the image and continuously condenses the image characteristics. The output of the last layer of the network contains rich information. The parameters obtained by training when inputting this information into the full link layer objectively reflect that some areas of the final feature map should be worthy of attention. By summing the results of all channels and inputting the attention heat map (Fig. 1), such a module is called an attention module. In this way, the net-

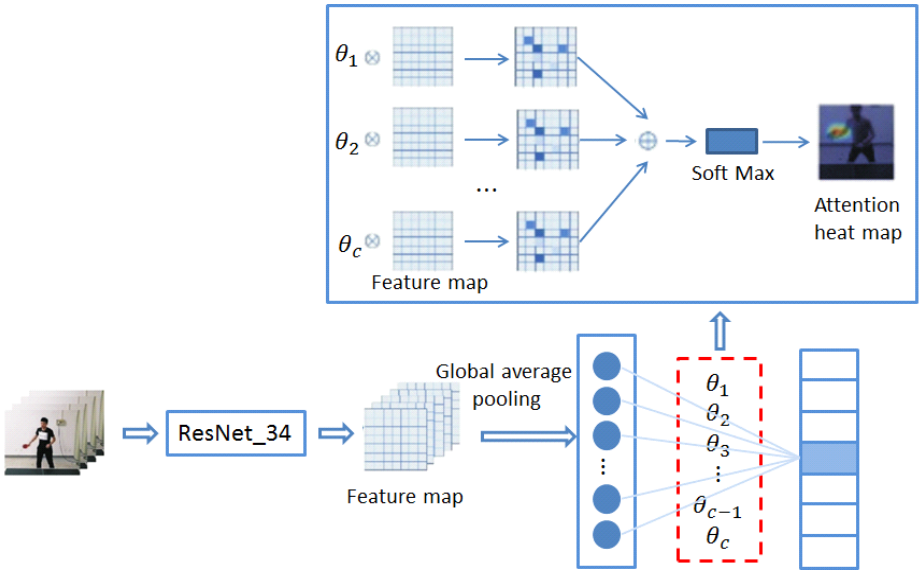

Fig. 1. Attention heat map generation.

work model can spontaneously focus on important parts, which is conducive to subsequent action classification.

3. PROPOSED METHOD

3.1 Action recognition based on skeleton data

Construct a spatial undirected graph structure on a skeleton sequence with N joint points and T frames $G=(V,E)$. Among them, the node collection $V=\{v_{ti}|t=1,\cdots,T,i=1,\cdots,N\}$ includes all nodes of the skeleton sequence. As the network input, the feature vector $X(v_{ti})$ on node v_{ti} is composed of the three-dimensional coordinate vector of the i-th joint point in the t-th frame. According to the connectivity of the human body structure, the joints in a frame are linked by edges, and each joint will be linked to the same joint point in consecutive frames. The edge set E consists of two subsets: the first subset $E_S=\{v_{ti}v_{tj}|(i,j)\in H\}$ represents the bone links of each frame, where H is the total number of human joint points; the second subset $E_F=\{v_{ti}v_{(t+1)i}\}$ represents the joint points connected in consecutive frames. As shown in Fig. 2.

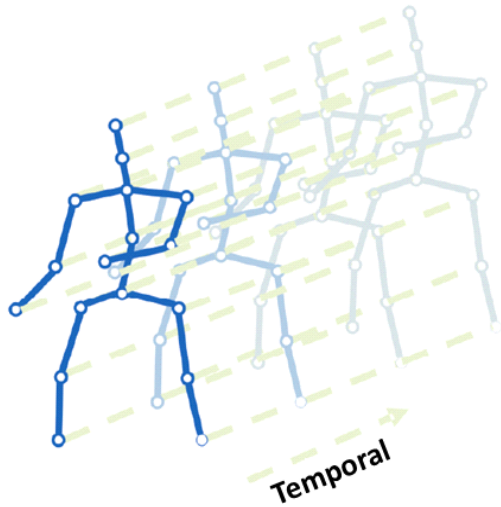

Fig. 2. Visual Structure of input data.

For the input data, first use graph convolution to encode the positional relationship of the joint points in space. In a single-frame human body joint point connection structure, a sampling function is defined on the neighbor set of the joint point v_{ti}; $B(v_{ti})=\{v_{tj}|d(v_{tj},v_{ti})\le D,i,j\in H\}$. Among them, $d(v_{tj},v_{ti})$ represents the minimum length of any path from v_{tj} to v_{ti}. Set D=1, that is, only the

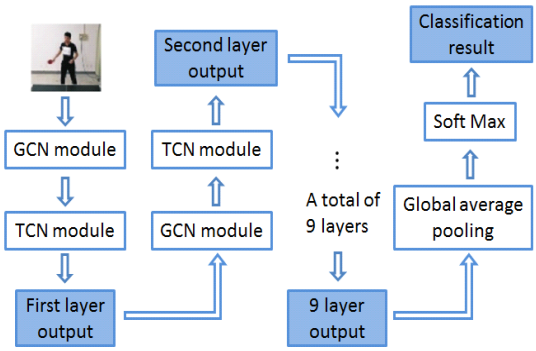

Fig. 3. Human Action recognition network based on skeleton data.

first-order neighbor nodes of node v_{ti} are selected. A single node v_{ti} of a single frame of video is aggregated by graph convolution and the feature output $F_{v_{ti}}$ can be expressed as:

$$F_{v_{ti}} = \sum_{j=1}^{N} D_{v_{ti}} A_{v_{ti}} x_{v_{ti}} \omega \tag{1}$$

Among them, $D_{v_{ti}}$ represents the degree matrix of the graph structure after sampling with the joint point v_{ti} as the central node, $A_{v_{ti}}$ represents the adjacency matrix of the graph structure after sampling with the joint v_{ti} as the central node, and ω represents the network parameters. The module for calculating the spatial structure of the joint points is called the GCN module.

It is relatively easy to connect the joints in the time dimension. In this paper, the time dimension is sampled at exactly the same time interval, so the joint connection of the time dimension is regular and can be calculated according to the traditional convolution method.Such modules are called TCN modules.

The method proposed in this paper alternately processes the two different dimensions of space and time. The network structure diagram is shown in Fig. 3.

3.2 Action recognition based on video data

The action recognition method based on video data will be affected by the video background information in the recognition effect. Therefore, this article first processes the input video data, manually crops out the areas where people are, and inputs the pictures of these areas as data into the recognition network.

The processed video data is input into ResNet in time series, the color information and contour information contained in the video frame are encoded through ResNet, and the encoded data with the attention mechanism is input into conv LSTM. The network structure is shown in Fig. 4.

3.3 Action recognition method based on dual-stream fusion

The dual-stream fusion method based on video data first extracts the dense optical flow of the video data, and uses CNN to process the video data and the optical flow data at the same time. The feature information encoded with CNN is input to the SoftMax classifier, which classifies each set of

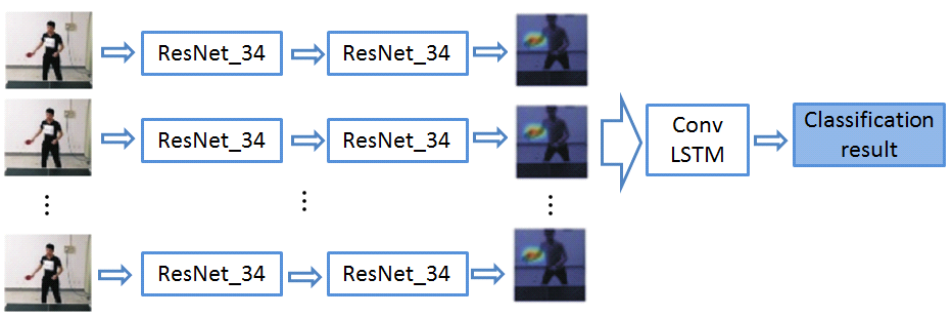

Fig. 4. Human Action recognition network based on video data.

data for testing, and enters the SoftMax classifier, which performs banditry on each set of data for testing Class, and output the probability that the set of data belongs to a certain class. The output probabilities of the two data streams are averaged to achieve information fusion. The comprehensive recognition rate after information fusion is often higher than the recognition rate of a single data stream.

This paper uses the human action recognition network under the spatio-temporal graph convolution framework and the human action recognition network based on video data, and trains the network. Use the converged parameters to test, and save the classification probability of each video segment in matrix form. After averaging the output probabilities of the two data streams, the final classification result is output, as shown in Fig. 5.

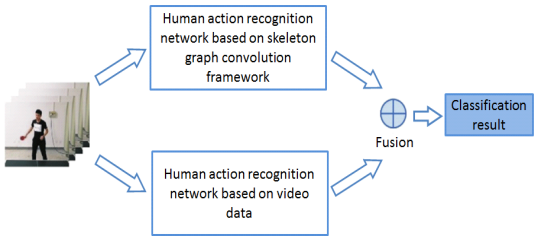

Fig. 5. Two-Stream fusion method.

4. EXPERIMENT AND ANALYSIS

The Kinetics human action dataset, The data set has 400 categories, each action has 400-1150 video clips, and the duration of each video is about 10 seconds. The current version has 306245 videos, divided into three parts, 250-1000 videos for each class during training, 50 videos for each class during verification, and 100 videos for each class during testing.

This data set only provides original video clips without skeleton data. Therefore, the original frame needs to be converted into a joint position sequence, and OpenPose is mainly used to capture the joint position. In order to obtain the joint positions, the video resolution was changed to 340 and the frame rate was converted to 30FPS. In this paper, the recognition performance is evaluated according to the classification accuracy of Top-1 and Top-5, and the data set is divided into a training set of 240,000 clips and a validation set of 20,000 clips.

The NTU RGB+D dataset is used to test the effectiveness of the method proposed in this paper. The data set contains 60 types of actions, a total of 56880 samples, of which 40 categories are daily behavior actions, 9 categories are health-related actions, and 11 categories are two-person interactions. These actions were performed by 40 people aged from 10 to 35 years old. This data set is collected by Microsoft Kinect v2 sensor, and uses three different angle cameras. The collected data forms include depth information, 3D bone information, RGB frames and infrared sequence.

There are two benchmarks for this data set. 1. Cross-Subject(CS): A training set of 40,320 clips and a test set of 16,560 clips. Under this division, some videos only appear in the training set, and some only appear in the test set. 2. Cross-View (CV): A verification set of 37,920 clips and a test set of 18,960 clips. Under this division, the clips used for training come from cameras 2 and 3, and the clips for the test set come from camera 1.

Based on the action recognition network of skeleton data, the initial learning rate is set to 0.1, and the learning rate is reduced to 10% every 10 epochs. The network is trained for a total of 80 epochs, and the standard cross-entropy loss function is used for back propagation.

After the completion of the action recognition based on video data, the initial learning rate is set to 0.001. The network is trained for 300 epochs, and the standard cross-entropy loss function is used for back propagation. The recognition rate is shown in Table 1.

On the Kinetics and NTU RGB+D datasets, a total of four skeleton-based action recognition

Table 1. Kinetics and NTU RGB+D data set recognition results.

Methods	Kinetics		NTU-RGB+D	
	Top-1/%	Top-5/%	CS/%	CS/%
Res-TCN[10]	0.2027	0.4045	-	-
STA-LSTM[11]	-	-	0.7338	0.8116
ST-GCN[12]	0.3074	0.5278	0.8154	0.8832
AS-GCN[4]	0.3483	0.5647	0.8672	0.9419
Proposed Method	0.3572	0.5729	0.8782	0.9233

methods are compared. RES-TCN(Residue Temporal Convolutional Networks) improves the interpretability of the model by reconstructing the spatio-temporal convolution with residual connections. STA-LSTM is a spatiotemporal attention LSTM network model established on the basis of a recurrent neural network with long and short-term memory. It can selectively focus on the joint differences of input frames and give different degrees of attention to the output of different frames, so it can extract distinguishable the temporal and spatial characteristics of sex help action recognition. ST-GCN breaks through the limitations of previous stock price modeling methods, applying graph convolution to human skeleton action recognition, and the proposed model has strong generalization capabilities. By combining A-LINKS and S-Link into a generalized skeleton diagram, AS-GCN further builds a behavior structure diagram convolutional network model, learns spatial and temporal characteristics, and can more accurately capture different action patterns.

According to the analysis of the experimental results in Table 1, the following conclusions can be drawn. 1. Compared with the Res-TCN model, graph convolution is more suitable for skeleton-based action recognition than traditional convolutional networks. 2. Compared with other network models based on LSTM (STA-LST, and AS-GCN), the method proposed in this paper can not only effectively capture skeleton data but also has superior recognition performance.

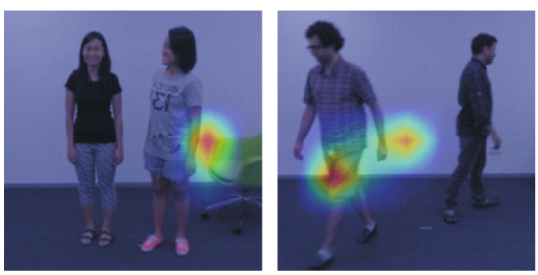

Fig. 6. Multi-person action attention heat map visualization.

But It can be found through the visualization results of the attention heat map of multi-person actions, as shown in Fig. 6. For the two-person actions in the data set, the attention heat map often only pays attention to one of them, while not paying enough attention to the other, which leads to the poor recognition ability of the recognition method, which affects the final recognition accuracy.

5. CONCLUSION

This paper proposes a dual-stream fusion method combining skeleton data and video data for action recognition. This method establishes a network model on the skeleton data and video data of human actions, and merges the output probabilities of the two network classifiers, which effectively realizes the information fusion of skeleton data and video data, and improves the recognition rate of human actions. In the next study, in order to make up for the lack of focus on only one person in the two-person action in the data set, the next step will be to propose a better information fusion method to give full play to the advantages of dif-

ferent types of data, achieve information complementation, and further improve the recognition rate.

REFERENCE

[1] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Proceeding of Adv. Neural Inf. Syst*, pp. 568-576, 2014.

[2] D.A. Huang and K.M. Kitani, "Action-Reaction: Forecasting the Dynamics of Human Interaction," *Proceeding of European Coference on Computer Vision*, pp. 489-504, 2014.

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

[4] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data," *Proceeding of The Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4263-4270, 2017.

[5] H. Zeyuan, P. Sangyun, and L. Eungjoo, "Human Motion Recognition Based on Spatio-Temporal Convolutional Neural Network," *Journal of Korea Multimedia Society*, Vol. 23, No. 8, pp. 977-985. 2020.

[6] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral Networks and Locally Connected Networks on Graphs," *Journal of Computer Vision and Pattern Recognition*, Online. https:// arxiv.org/abs/1312.6203. 2013.

[7] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Ltering," *Journal of Neural Information Processing Systems*, pp. 3844-3852, 2016.

[8] T.N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *Journal of Machine Learning*, Online. Available:http://arxiv.org/abs/1609.02907, 2016.

[9] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning Convolutional Neural Networks for Graphs," *Proceeding of The 33nd International Conference on Machine Learning*, pp. 2014-2023, 2016.

[10] T. Kim and A. Reiter, "Interpretable 3D Human Action Analysis with Temporal Convolutional Networks," *Proceeding of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1623-1631, 2017.

[11] R. Li, M. Tapaswi, and R. Liao, et al, "Situation Recognition with Graph Neural Network," *Proceeding of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4173-4182, 2017.

[12] M. Simonovsky and N. Komodakis, "Dynamic Edge-Conditioned Filters in Convolutional Neural Network on Graphs," *Proceeding of The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3693-3702, 2017.

Zeyuan Hu

was born in Dalian, Liaoning. P.R. China, in 1992. He received the bachelor's degree in Automation from Qingdao Institute of Technology, P.R. China(2011-2015). He received the master's degree in Information Communication Engineering from Tongmyong University, Busan, Korea(2016-2018). Currently, he has been studying for his doctoral degree in the Department of Information and Communications Engineering in Tongmyong University, Korea. And he is majoring in image processing and pattern recognition.

Yiran Feng

He is a doctoral candidate majoring in electronic information and Communication at Tongmyong University, South Korea. He received the B.E. degree and M.E. degree from Dalian Polytechnic University, CHINA in 2013 and 2015 respectively. He works as a teacher at Dalian Polytechnic University and member of Korea Multimedia Association. He is mainly engaged in image recognition and robot technology research.

Lee Eung-Joo

received his B. S., M. S. and Ph. D. in Electronic Engineering from Kyungpook National University, Korea, in 1990, 1992, and Aug. 1996, respectively. Since 1997 he has worked with the Department of Information & Communications Engineering, Tongmyong University, Korea, where he is currently a professor. From 2005 to July 2006, he was a visiting professor in the Department of Computer and Information Engineering, Dalian Polytechnic University, and from Dec 2018 he was appointed honorary professor of Dalian Polytechnic University, China. His main research interests include biometrics, image processing, and computer vision.