Computational Support for Two-Person Interactive Behavior Recognition Based on Multi-Channel Spatio-Temporal Fusion Network Using Skeleton Joint Data

Yang Cao[1*]

[1]College of Electronic Information Engineering, Chongqing Technology and Business Institute, Chongqing 401520, China, yangcao20230501@163.com

Corresponding author: Yang Cao, yangcao20230501@163.com

Abstract. This paper presents a two-person interactive behavior recognition method based on a multichannel space-time fusion network of skeletal joint point data. First, the single and two-person behavioral characteristics of both sides of the interaction are represented by distance. Digitized single-person and two-person interactive behaviors are input into a multi-channel CNN-LSTM network in time order to learn the time and space characteristics. The multi-channel CNN-LSTM model is a two-level cascade network. The first layer of CNN network is responsible for learning spatial characteristics, and the second layer of LSTM network is responsible for learning time characteristics. Finally, the spatial and temporal characteristics learned are classified by the full connectivity layer and the Softmax layer. The method in this paper is tested on NTU RGB+D dataset, and the accuracy of cross-angle experiment can reach 90.3%. The experimental results show that the accuracy of the proposed method is better than the typical method in the field of two-person interactive behavior recognition.

Keywords: two-person Interactive behavior;distance;CNN-LSTM;spatiotemporal fusion network; multi-channel; Computational Support
DOI: https://doi.org/10.14733/cadaps.2024.S9.130-137

1 INTRODUCTION

With the rapid development of computer vision and machine learning technologies, utilizing video data for human behavior recognition and analysis has become a hot research topic in the fields of computer science and artificial intelligence. Video data, as a rich source of information, provides abundant spatial and temporal contextual cues, enabling more accurate and comprehensive understanding and analysis of human behaviors[16].

The application of video data in two-person interaction behavior has wide applications in fields such as human-computer interaction [4], virtual reality [14], and intelligent monitoring [18]. For

example, studying interaction behaviors based on video data can provide important references for interpersonal interaction and interactive design in virtual environments. By observing and analyzing behavioral patterns and regularities in two-person interactions, guidance can be provided for generating and designing behaviors of virtual characters, enhancing immersion and realism in virtual reality [2][7].

two-person interactive behavior recognition can be categorized into two types based on different data sources: RGB video data and skeletal joint data[11][22]. The recognition based on skeletal joint data involves acquiring the positions and dynamic information of human skeletal joints using sensors, depth cameras, or pose estimation algorithms[15]. Compared to RGB video data, skeletal joint data offers more stable and reliable data quality. It accurately reflects the movement status of the human body and avoids issues such as lighting conditions and background interference that can affect behavior recognition. Due to the advantages of skeletal joint data, research on two-person interactive behavior recognition based on joint data has become a hot topic.

Currently, two-person interactive behavior recognition based on joint data can be broadly categorized into traditional methods and deep learning methods. Traditional learning methods construct and extract features from joint data based on domain experts' knowledge and experience, and then feed these features into classifiers to predict the interactive behaviors. For example, Slama et al. [13] describe an action as a collection of three-dimensional joint coordinates in a time series. Each action sequence is represented as a linear dynamical system that generates three-dimensional joint trajectories. They use an autoregressive moving average model to represent the sequences and finally employ a linear support vector machine (SVM) for classification. Yun et al. [24] utilize the distances between all joint pairs of the two individuals in the current frame, the distances between joint pairs in the current frame and the previous frame, and the distances among individual joints in the current frame to represent the human pose. These representations are then fed into a support vector machine (SVM) for recognition.Traditional methods heavily rely on expert knowledge and experience, and they often suffer from limited transferability and generalization capabilities. Therefore, this paper intends to adopt a deep learning-based approach to automatically learn features, aiming to improve the transferability and generalization capabilities of the model. computational support, such as powerful hardware resources and efficient algorithms, the deep learning model can efficiently process and analyze large amounts of joint data. This computational support enhances the model's capacity to learn intricate features and improves its ability to generalize to unseen data.

The recognition of two-person interactive behavior based on deep learning primarily utilizes Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to automatically extract features. Ke et al. [9] propose a method for learning fragment representations by generating three segments corresponding to the three channels (x, y, z) of joint point coordinates. These segments are then fed into a CNN for recognition. [10] introduces a novel Independent Recurrent Neural Network (IndRNN) for two-person interactive behavior recognition. In this network, neurons within the same layer are independent of each other but connected across different layers, effectively preventing the issues of gradient explosion and vanishing. This enables the network to capture long-term dependencies.The aforementioned methods mainly focus on representing single-person behavior features while neglecting the representation of the interplay between the two individuals. Therefore, this paper designs a feature representation based on distance and utilizes a CNN-LSTM network to automatically learn features for two-person interactive behavior recognition.

2 FEATURE REPRESENTATION BASED ON DISTANCE

Distance-based feature representation is a common feature representation method used in joint point data behavior recognition. It has simple features and is not affected by scale changes. In this

paper, single person centerpoint Oa and Ob are selected for behavioral identifiers a and b, and the distances daiao and dbibo from all nodes of identifier a and b to centerpoint Oa and Ob are calculated respectively. Individual behavior characteristics are represented by dia and DIB values. The distance formulas for daiao and dbibo are shown in formulas 1 and 2 .

$$d_{aiao}=\sqrt{\left(X_{ai}-X_{ao}\right)^2+\left(Y_{ai}-Y_{ao}\right)^2+\left(Z_{ai}-Z_{ao}\right)^2} \quad i\in(1,2,3.....L) \tag{1}$$

$$d_{bibo}=\sqrt{\left(X_{bo}-X_{bo}\right)^2+\left(Y_{bo}-Y_{bo}\right)^2+\left(Z_{bo}-Z_{bo}\right)^2} \quad i\in(1,2,3....L) \tag{2}$$

Relative distances between two person's joint points in two-person interactive behavior recognition can represent the characteristics of two-person behavior association in two-person interactive behavior. The relative distance daibi of two persons a and b is shown in formula 3.

$$d_{aibi}=\sqrt{\left(X_{ai}-X_{bi}\right)^2+\left(Y_{ai}-Y_{bi}\right)^2+\left(Z_{ai}-Z_{bi}\right)^2} \quad i\in(1,2,3....L) \tag{3}$$

The association features Da, Db and Dab of single and two-person interaction behavior of joint data series can be obtained, as shown in formula 4.

$$\begin{cases} D_a=\left\{daiao(\mathrm{t}),t\in(1,2,3....T),i\in(1,2,3....L)\right\} \\ D_b=\left\{dbibo(\mathrm{t}),t\in(1,2,3....T),i\in(1,2,3....L)\right\} \\ D_{ab}=\left\{daibi(\mathrm{t}),t\in(1,2,3....T),i\in(1,2,3....L)\right\} \end{cases} \tag{4}$$

3 MULTICHANNEL SPATIAL-TEMPORAL FEATURE FUSION NETWORK

The network of multichannel space-time feature fusion consists of three channels C1, C2 and C3. C1 and C2 are respectively responsible for extracting the spatiotemporal information of the data series Da and Db of the interactive actor, while C3 is responsible for extracting the spatiotemporal information of the associated features of the interactive actor a and B. Each channel uses CNN-LSTM model to extract features. The CNN-LSTM network model consists of two cascade structures. The first layer extracts spatial features, the second layer extracts temporal features, the full connection layer outputs feature information containing spatial and temporal features, and the interactive behavior is classified through the softmax layer. A network of multichannel spatiotemporal feature fusion is shown in Figure 1.

The CNN network is responsible for extracting spatial features and consists of two convolution layers, pooling layers and Dropout layers. The convolution layer uses M and N convolution cores of 1D-CN N to convolute joint sequence data with input length T and dimension N. The formulas are shown in Formulas 5 and 6, where the output of two convolution layers and the weight matrix of two convolution cores are represented [8].

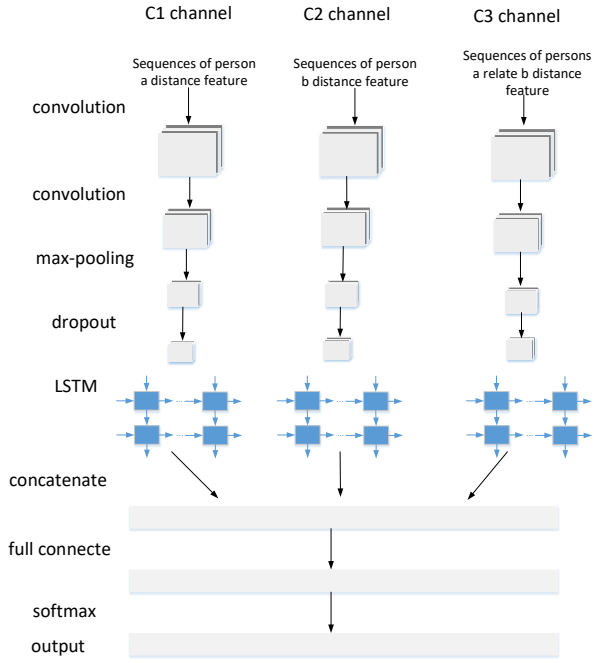

Figure 1: Multichannel Spatial-Temporal Feature Fusion Network.

$$\mathrm{f}_m = \sigma(w1_{(L,l)} * F + b), m = (1, 2, ..., M) \quad (5)$$

$$\mathrm{f}_s = \sigma(w2_{(M,l)} * f_m + b), s = (1, 2, ..., N) \quad (6)$$

$$\mathrm{f}_s = CNN(F) \quad (7)$$

The time-related characteristics of joint sequence data were studied using a long-term and short-term memory network (LSTM). The LSTM network definition is shown in formula 8. Among them are the forgetting gate, the input gate, the output gate, the new candidate value, the memory gate, and the output value. LSTM network avoids the disappearance of gradients caused by the recurrent neural network (RNN)[19][5].

$$\begin{aligned} f_t &= \sigma(W_f.[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i.[h_{t-1}, x_t] + b_i) \\ o_t &= \sigma(W_o.[h_{t-1}, x_t] + b_o) \\ \tilde{C}_t &= \tanh(W_c.[h_{t-1}, x_t] + b_c) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ h_t &= o_t * \tanh(C_t) \end{aligned} \quad (8)$$

Input the spatial characteristics of the CNN network output into the LSTM network, learn the time-related characteristics of the joint point sequence data, and output the space-time eigenvector of the feature, as shown in formula 9.

$$f_{st} = LSTM(f_s) \tag{9}$$

Three sets of Da, Db and Dab sequences are input into the network of multichannel space-time feature fusion, and the CNN-LSTM network model is represented by a function. The features of each channel space-time fusion are shown in formula 10.

$$\begin{aligned} f_{std_a} &= F_{st}(D_a) \\ f_{std_b} &= F_{st}(D_b) \\ fstd_{ab} &= F_{st}(D_{ab}) \end{aligned} \tag{10}$$

The spatial-temporal fusion characteristics of the output from each channel are connected side by side as shown in formula 11.

$$f_{fusion} = [f_{std_a}, f_{std_b}, f_{stda_b}] \tag{11}$$

Output the probability of identifying categories for two-person interaction using the full connection layer and the softmax function, as shown in formula 12.

$$O = soft\max(W * f_{fusion}) \tag{12}$$

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Dataset Introduction

NTU RGB+D dataset is a large-scale RGB-D dataset for human motion recognition and analysis[12]. It combines RGB color images with depth images to capture rich spatial and temporal information and provide a more comprehensive and accurate description of human motion.The dataset uses multiple RGB cameras and depth sensors for data collection, including three data sources: RGB, depth and skeleton data. This paper evaluates the performance of the model by choosing the common two-person interaction behaviors of skeleton data in NTU RGB+D dataset, such as hugging, pushing, hitting, and carrying items.

4.2 Experimental Settings

The experiment was performed on a Ubuntu 16.04 system using a RTX2070 graphics card with Tensor Flow as the backend Keras framework. The learning rate is initialized to 0.005. After each 4 rounds, the learning rate is reduced to one tenth of the original learning rate, the maximum number of iterations is 16, the batch size is 12, the network is optimized by rmsprop, and the loss function is cross-entropy.There are two types of protocol for the official evaluation methods of NTU RGB+D datasets: cross-subject (CS) and cross-view (CV)[12]. In order to cope with the situation with different and variable perspectives in the actual scene, this paper uses the Cross-perspective CV protocol to evaluate.

4.3 Experimental and test results

Method	Accuracy/%
Lie Group	56.7
TS-LSTM	78.6
Trust Gate ST-LSTM	79.2
TSRJI -CNN	82.3
AGC-LSTM	88.7
Our method	90.3

Table 1: Accuracy for Human Action Recognition for NTU-RGBD Dataset.

From Table 1, it can be seen that the accuracy of model recognition proposed in this paper is due to traditional methods and methods that only use CNN or LSTM models. In this paper, the proposed model achieves 90.3% recognition accuracy in cross-angle experiments.The accuracy of 11 types of two-person interaction for NTU-RGB+D datasets is shown in Figure 2 using the model proposed in this paper. The recognition results of all behavior categories were above 80%.

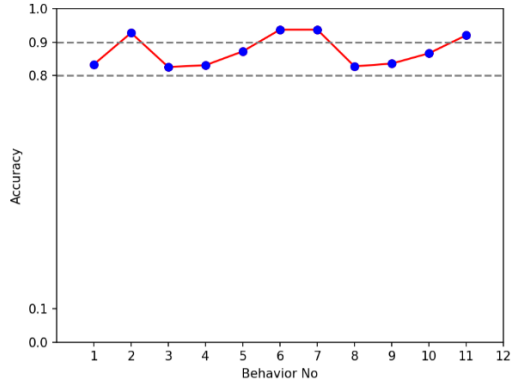

Figure 2: Accuracy of Cross-View Experiments.

5 CONCLUSION

This paper presents a method for identifying two-person interactive behaviors based on skeletal joint point data. The single-person and two-person interactive behaviors of the joint point data are characterized by distance, a multi-channel space-time feature fusion network is designed using CNN-LSTM model to extract the behavior features, and the behavior types are judged using full connectivity and softmax layers. The method proposed in this paper has high accuracy in identifying two-person interactive behaviors. The RGB video features will be introduced in the future, and the behavior recognition analysis in complex scenes will be achieved through the fusion of multimodal features.

Yang Cao, https://orcid.org/0009-0001-1952-343X

ACKNOWLEDGEMENT

Supported by the Science and Technology Research Program of Chongqing Municipal Education Commission（Grant NO. KJQN202204008.

REFERENCES

[1] Abdulazeem, Y.; Balaha, H. M.; Bahgat, W. M.; Badawy, M.: Human Action Recognition Based on Transfer Learning Approach, IEEE Access, (9), 2021, 1-1. https://doi.org/10.1109/ACCESS.2021.3086668

[2] Almeida Leonardo, T.: et al. A Video-Based Approach for Real-Time Recognition of Body Gestures in Virtual Reality Applications, Sensors, 20(12), 2020, 3502. https://doi.org/10.3390/s20123502

[3] Caetano, C.; Bremond, F.; Schwatz, W.R.: Skeleton Image Representation for 3d Action Recognition Based on Tree Structure and Reference Joints, 32nd Sibgrapi Conference on Graphic Patterns and Images, 2019, 16-23. https://doi.org/10.1109/SIBGRAPI.2019.00011

[4] Chen, C.: et al. Real-Time Human-Robot Collaboration Recognition Based on Hierarchical Attention Mechanism, IEEE Transactions on Industrial Electronics, v68, (10), 2021, 8243-8253.

[5] Chen, X.; Beike, Z.; Dong, G.: Bearing Fault Diagnosis Base on Multi-Scale CNN and LSTM Model, Journal of Intelligent Manufacturing, 32(4), 2021, 971-987. https://doi.org/10.1007/s10845-020-01600-2

[6] Dwyer, T.; Tien, P. W.; Wei, S.; Calautit, J. K.; Darkwa, J.; Wood, C. : Vision-Based Human Activity Recognition for Reducing Building Energy Demand:. Building Services Engineering Research & Technology, 42(6), 2021, 691-713, https://doi.org/10.1177/01436244211026120

[7] Gupta, P.: et al. Real-Time Hand Pose Estimation using Depth Sensors, Multimedia Tools and Applications, 76, (20), 2017,21219-21241. https://doi.org/10.1007/s11042-017-4961-6.

[8] Ha, S.; Jeong-Min Y.; Seungjin C.: Multi-modal convolutional Neural Networks for Activity Recognition, 2015 IEEE International conference on systems, Man, and Cybernetics. IEEE, 2015. https://doi.org/10.1109/SMC.2015.525

[9] Ke, Q.; Bennamoun M.; An, S.: et al. A New Representation of Skeleton Sequences for 3d Action Recognition, IEEE Conference on Computer Vision and Pattern Recognition, 2017, 3288-3297. https://doi.org/10.1109/CVPR.2017.486

[10] Li, S.;Li, W.;Cook, C.:et al. Independently Recurrent Neural Network (indrnn) : Building a Longer and Deeper Rnn, IEEE Conference on Computer Vision and Pattern Recognition, 2018, 5457-5466. https://doi.org/10.1109/CVPR.2018.00572

[11] Poppe, R. A Survey on Vision-Based Human Action Recognition, Image and vision computing 28(6), 2010, 976-990. https://doi.org/10.1016/j.imavis.2009.11.014

[12] Shahroudy, A.;Liu, J;Ng, T. T. :et al. Ntu rgb+ d: A Large Scale Dataset for 3d Human Activity Analysis, IEEE Conference on Computer Vision and Pattern Recognition, 2016, 1010-1019, https://doi.org/10.1109/CVPR.2016.115

[13] Slama, R.; Wannous, H.; Daoudi, M.: et al. Accurate 3D Action Recognition Using Learning on the Grassmann Manifold, Pattern Recognition, 48(2), 201, 556-567. https://doi.org/10.1016/j.patcog.2014.08.011

[14] Sridharan, S.: et al. Real-Time Gesture Recognition for Virtual Reality Applications using Deep Learning, Proceedings of the 2018 International Conference on Virtual Rehabilitation, IEEE, 2018, 1-2. https://doi.org/10.1109/ICVR.2018.8465974.

[15] Tsai, M. F. ; Chen, C. H. : Spatial Temporal Variation Graph convolutional Networks (stv-gcn) for Skeleton-Based Emotional Action Recognition, IEEE Access, 9, 2021, 13870-13877, https://doi.org/10.1109/ACCESS.2021.3052246

[16] Wang, P.;Li, W.;Ogunbona, P. et al. RGB-D-Based Human Motion Recognition with Deep Learning: A Survey. Computer Vision and Image Understanding, 171, 2017, 118-139. https://doi.org/10.1016/j.cviu.2018.04.007
[17] Wang,P. Li,Z. Hou,Y. :et al. Action Recognition Based on Joint Trajectory Maps Using convolutional neural networks, 24th ACM international conference on Multimedia, 2016, 102-106. https://doi.org/10.1145/2964284.2967191
[18] Wu, T.: et al. Real-Time Abnormal Human Behavior Recognition for Intelligent Video Surveillance, Sensors, 20, 7, 2020, 1961.
[19] Xia, K. Jianguang, H.; Hanyu, W.: LSTM-CNN architecture for human activity recognition, EEE Access 8 2020, 56855-56866. https://doi.org/10.1109/ACCESS.2020.2982225
[20] Yahaya, S. W.; Lotfi, A.; Mahmud, M.; Towards a Data-Driven Adaptive Anomaly Detection system for Human Activit, Pattern Recognition Letters, 145(6), 2021, https://doi.org/10.1016/j.patrec.2021.02.006
[21] Yamini, G.; Ganapathy, G.: Enhanced sensing and Activity Recognition System Using Iot for Healthcare, International Journal of Information Communication Technologies and Human Development, 13(2), 2021 , 42-49. https://doi.org/10.4018/IJICTHD.2021040103
[22] Ye, Q. ; Li, R. ; Guo, X. : Human Interaction Behavior Recognition Based on Local Spatial - Temporal and Global Feature, Journal of Physics: Conference Series, 1754(1), 2021, 012188 (6pp). https://doi.org/10.1088/1742-6596/1754/1/012188
[23] Ye, Q. ; Li, R. ; Yang, H. ; Guo, X: Human Interactive Behaviour Recognition Method Based on Multi-Feature Fusion, International Journal of Computational Science and Engineering, (3), 2022, 25. https://doi.org/10.1504/IJCSE.2022.123113
[24] Yun, K. ; Honorio, J. Chattopadhyay, D. et al. Two-Person Interaction Detection Using Body-Pose Features and Multiple Instance Learning, Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2012, 28-35. https://doi.org/10.1109/CVPRW.2012.6239234
[25] Zhang, C.; Li, M.; Wu, D. Federated Multidomain Learning With Graph Ensemble Autoencoder GMM for Emotion Recognition, IEEE Trans. Intell. Transp. Syst. 2022; Early Access. https://doi.org/10.1109/TITS.2022.3203800