Comparative study on neural network models of human behavior recognition methods

Li Jiazhi[*1,*2], Liu Ning[*1,*2], Zhang Dedi[*1,*2] , Liu Fuchao[*1,*2]

[*1]Beijing Information Science and Technology University, Technology Beijing Key Laboratory of High Dynamic Navigation Technology, Beijing 100192, China
E-mail: ning.liu@bistu.edu.cn
[*2]Key Laboratory of Modern Measurement and Control, Beijing 100101, China
E-mail: ning.liu@bistu.edu.cn

To solve the problem of human behavior feature extraction and classification, a human behavior recognition and classification method based on convolutional neural network (CNN) and long short-term memory network (LSTM) was proposed .Due to the strong correlation between the front and back movements of the human body, compared with the single convolutional neural network model, the model integrated with the long and short memory network can more accurately identify the six movements of walking, jogging, going upstairs, coming down, standing and sitting. The convolutional neural network is used to extract the acceleration features, and then the extracted features are input to the long and short term memory network for classification, to complete the human behavior recognition. Through simulation and verification, the CNN-LSTM hybrid model can achieve an average accuracy of 96.9% for action recognition by testing in the open WISDM data set and observing the results of several experiments, which verifies the feasibility of this method.

Keywords: Convolutional Neural Network, accelerometer, Human behavior recognition, wearable devices, Long short-term memory network

1. INTRODUCTION

In recent years, with the development of information technology and wearable devices, the research of human behavior recognition technology has attracted much attention and has gradually become a major research hotspot in the field of machine vision. The state recognition of the human body is mainly through analysis of data collected by sensors such as accelerometers and gyroscopes, which can identify the movement characteristics and behavior patterns of the human body. At present, human behavior recognition technology is widely used in individual soldier equipment, fire rescue, rehabilitation training, outdoor sports, intelligent machines and many other aspects, and it has broad market prospects and economic value.[1]

Traditional human behavior recognition methods include support vector machines, naive Bayes, K-nearness, decision trees, etc. The key point is behavior representation and behavior recognition. Behavior representation mainly includes feature formation, feature selection, feature extraction, and feature extraction. The purpose is to extract distinguishable behaviors and classifiable action features[2-4]. The purpose of behavior recognition is to recognize actions. The accuracy of these traditional methods for gesture recognition depends on the choice of classifier and the use of sensors, and the extraction of features mainly relies on manual labor, which leads to huge workload and heavy tasks[5,6]. With the development of convolutional neural networks, more and more scholars have begun to notice the superiority of neural networks for image processing and the convenience of automatically extracting features, which have natural advantages in the processing of human body state recognition.[7-10]

Among the training models of neural networks, there are also deep neural networks, convolutional neural networks, long and short-term memory networks, and so on. This paper studies and compares the recognition accuracy of the six daily motion states of walking, jogging, going upstairs, going downstairs, standing, and sitting under different neural network training. The final result shows that the accuracy of the neural network under the combined model is better than that of a single network. The improvement of the accuracy of action recognition effectively guarantees the personal safety of firefighters during rescue, improves the comfort assessment of prosthetic wearers, and accurately locates during outdoor sports. The results show that the combined neural network has better performance in the field of human state recognition. Broad market prospects and development value. [10-13]

2. NETWORK MODEL

2.1. Deep Neural Networks

A Deep Neural Networks is composed of an input layer, a hidden layer, and an output layer. Compared with ordinary neural networks, the depth of the network is more emphasized. The hidden layer can be composed of multiple layers, and each layer is connected in a fully

connected manner. For the i-1 layer and the i layer, each node in the i-1 layer is connected to all the nodes in the i layer, that is, when each node in the nth layer is performing the calculation, the input activation The function is the weighting of all nodes in the n-1 layer. And each layer uses the output of the previous layer as input, and then goes to the next layer after calculation.

$$h_1 = A_1(W_1 x + b_1) \tag{1}$$

$$h_i = A_i(W_i h_{i-1} + b_i), 1 < i < d \tag{2}$$

Where h_i is the output value of the i-th hidden layer; W_i is the weight value; b_i is the bias; A_i is the activation function corresponding to the hidden layer; d is the depth of the network, that is, the sum of the number of hidden and output layers.

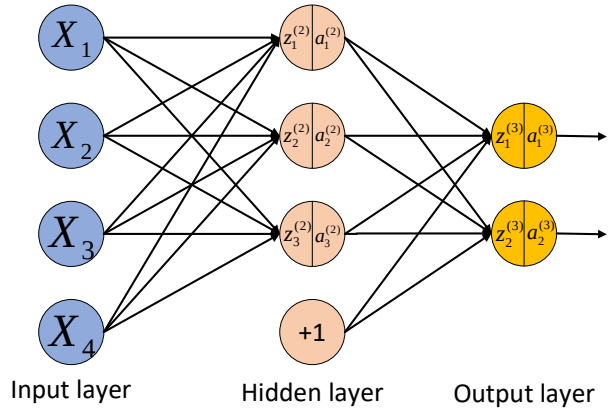

Figure 1 Fully connected network diagram

2.2. CONVOLUTIONAL NEURAL NETWORK

The excellent classification effect of convolutional neural networks in the field of image recognition has promoted the development of convolutional neural networks in other fields such as HAR. In the human body state recognition based on accelerometer sensors, scholars have found that the use of neural networks can achieve better Performance. The x, y, and z-axis three-dimensional data measured by the accelerometer are equivalent to the three RGB channels in image recognition. By identifying the graphic features of the three-axis accelerometer, feature extraction and feature comparison are performed to identify the human body's motion state. Convolutional neural networks have been applied in many fields due to their excellent functions such as pooling operations, local connections, weight sharing, and multi-layer structure. For huge human dynamic behavior data sets, trained convolutional neural networks can have the same behavior Classification of behaviors with characteristics but some local differences to achieve the effect of identifying behaviors. The product neural network model in this paper is constructed by Sequential. The architecture of a convolutional neural network is mainly composed of an input layer, a convolutional layer, a pooling layer, a fully connected layer, and a Softmax layer.

Input layer: The input layer is the input of the entire neural network. The input data in the human body gesture recognition is the x, y, and z-axis three-dimensional data measured by the pre-processed accelerometer, which is equivalent to the three RGB channels in image recognition. Before the data is input, preprocessing is required to reduce the training time and speed up the convergence of the network. Common processing methods include de-averaging, normalization, and SVD dimensionality reduction methods.

Convolutional layer: The convolutional layer is the core of building a convolutional neural network. Its main function is to extract the features of the image, and conduct a deeper analysis of each piece of content in the neural network to obtain features with a higher degree of abstraction.

$$x_j^{l+1} = \sum_{i=1}^{n} x_i^{(l)} * k_j^{(l+1)} + b_j^{(l+1)} \tag{3}$$

Where x_j^{l+1} represents the j-th channel feature output by the l+1 layer, $k_j^{(l+1)}$ represents the j-th convolution kernel of the feature map of the i-th channel from the previous layer to the current layer, and $x_i^{(l)}$ represents the feature value of the i-th channel output by the l layer , $b_j^{(l+1)}$ represents the offset of the convolution kernel of the j channels in the l+1th layer.

Pooling layer: The pooling layer is located between continuous convolutional layers. The pooling layer can further reduce the parameter nodes and compressed data in the fully connected layer, reduce the degree of overfitting, but will not change the original features. It is a way of dimensionality reduction. Common pooling methods include mean pooling and maximum pooling. The average pooling method is to take the average value in the corresponding area of each pooling window as the pooled value. The maximum pooling method is to take the maximum value in the corresponding area of each pooling window as the pooled value..

Fully connected layer: After multiple rounds of convolution and pooling processing, at the end of the convolutional neural network, 1-2 fully connected layers will generally give the final classification results, and the extracted ones contain higher information content In the end, a fully connected layer is required to complete the classification results.

Softmax layer: The Softmax layer is mainly used for classification problems. After processing, the probability distribution of different types in the current sample can be obtained to achieve more accurate classification.

$$s_i = \frac{e^{Z_i}}{\sum_K e^{Z_k}} \tag{4}$$

$$z_i = \sum_j x_{ij}.w_{ij} + b \tag{5}$$

Among them, s_i represents the output value of the i-th neuron, w_{ij} is the j-th weight of the i-th neuron, and b is the bias value. z_i represents the i-th output of the network.

Over-fitting treatment: Over-fitting is a training mode with complete memory. When a new state appears, it will often lead to misjudgment. The main performance is that the error of the model on the training set is small, and the error on the test set is large. When the model is trained, the dropout layer will place hidden neuron nodes in an inactive state according to a certain random probability, reducing the interaction and complex co-adaptability between hidden neuron nodes. During training It will reduce the over-fitting phenomenon, increase the robustness, and improve the generalization vitality.

2.3. LONG SHORT-TERM MEMORY NETWORK

LSTM is an extension of RNN (recurrent neural network). Recurrent neural network can effectively deal with time series problems. However, as the training model deepens and the time series becomes longer, RNN no longer meets the requirements. Need to introduce LSTM to supplement and solve this problem. In daily life, people's actions are continuous and cannot jump. For example, the current action is jogging, then the previous action and the next action are also highly likely to be jogging. Therefore, in the recognition of human behavior, simply relying on the three-axis acceleration data of the accelerometer to extract features for classification will produce a certain error, and LSTM will make up for this shortcoming, taking into account the sequence of actions before and after, and long-term storage memory.

The unit structure of LSTM includes input gates, forget gates and output gates, of which the forget gate plays a key role. The function of the input gate is to control the addition of new information flow, the output gate is used to control the value of the output, and determine the next hidden state, and the function of the forget gate is to decide which information to discard or retain. The following is its parameter formula and unit structure diagram.

$$f_t = \sigma(w_{xf} x_t + w_{hf} h_{t-1} + w_{cf} c_{t-1} + b_f) \quad (6)$$

$$i_t = \sigma(w_{xi} x_t + w_{hi} h_{t-1} + w_{ci} c_{t-1} + b_i) \quad (7)$$

$$o_t = \sigma(w_{xo} x_t + w_{ho} h_{t-1} + w_{co} c_{t-1} + b_o) \quad (8)$$

$$c_t = f_t . c_{t-1} + i_t . \tanh(w_{xc} x_t + w_{hc} h_{t-1} + b_c) \quad (9)$$

$$h_t = o_t . \tanh(c_t) \quad (10)$$

Among them, c_t represents the state of the memory cell at time t, h_t represents the output of the hidden layer, and b_f 、 b_i 、 b_o and b_c represent the bias. The parameters $w_{x(f/i/o)}$ 、 $w_{h(f/i/o)}$ and $w_{c(f/i/o)}$ represent the weight value. The weight value needs to be obtained through training.

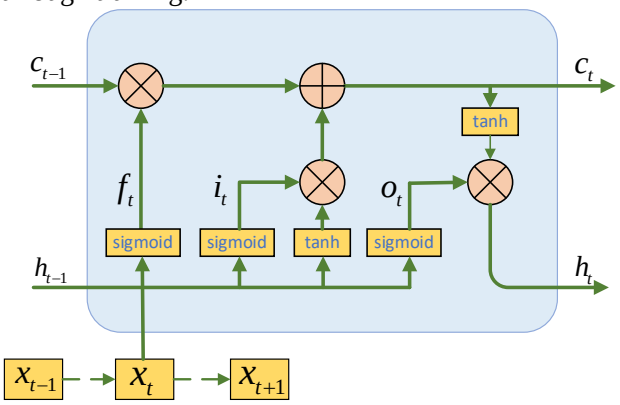

Figure 2 LSTM unit structure diagram

2.4. CNN-LSTM hybrid model

This model integrates CNN and LSTM, and makes full use of their respective advantages to complete human behavior recognition. The advantage of CNN is that it can efficiently perform feature extraction and map the extracted features to feature maps, while the advantage of LSTM is to build a long- and short-term memory network model, which can be more accurate through the state of the previous moment and the next moment. Determine the current state, and then send the result to the Softmax layer for final classification.

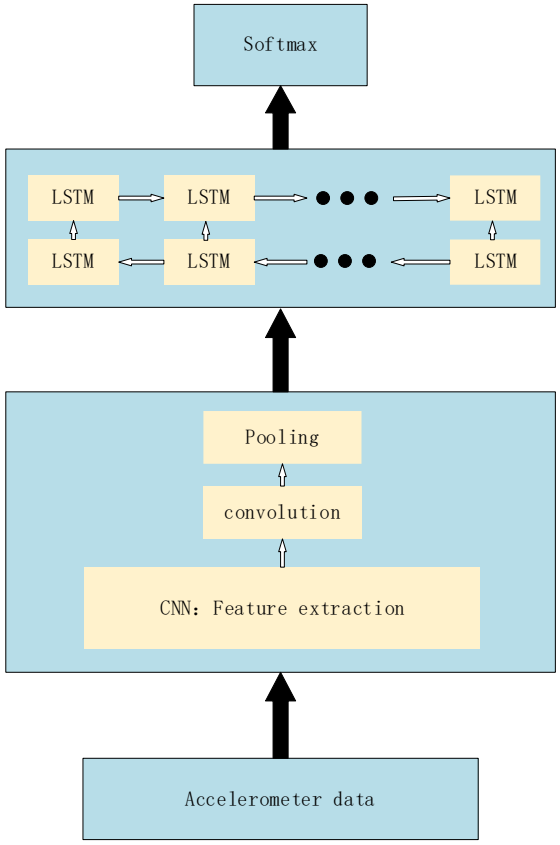

Figure 3 CNN-LSTM network model

3. TEST AND RESULT ANALYSIS

3.1. Data set

The data set used in the model in this article is the WISDM public data set. The sensor is an accelerometer sensor built into the mobile phone. The sampling frequency is 20HZ, which means 20 data per second. The data set is randomly divided into 80% training set and 20% test set for neural network model training and state classification. The data set includes six actions: walking, jogging, going upstairs, going downstairs, standing, and sitting down. Each row of the data set represents a data sample, and each data sample is the data of the x, y, and z three-axis measured by the accelerometer. , The total sample size is 1098203, of which the walking sample is 424397, accounting for 38.6% of the total sample; the jogging sample is 342176, accounting for 31.2% of the total sample; the upstairs sample is 122869, accounting for 11.2% of the total sample; the downstairs sample is 100427. Accounted for 9.1% of the total sample; the sitting sample was 59,939, accounting for 5.5% of the total sample; the standing sample was 48,395, which accounted for 4.4% of the total sample.

Because the convolutional neural network is very sensitive to the imbalance of the categories in the training samples. Compared with the unbalanced training samples, the balanced training samples can achieve better results. Therefore, the sample balance should be carried out before training. The sample size is the minimum of the above six sample sizes, that is, the number of samples in each group is 48,395, and the total number of samples is 290,370.

In the sample of the overall data set, 500 sets of each action are taken, and a total of 3000 sets of acceleration data are analyzed. Among them, 0-500 is jogging; 501-1000 is walking, 1001-1500 is upstairs, 1501-2000 is downstairs, 2001-2500 is sitting down, and 2501-3000 is standing. Analyze its acceleration data.

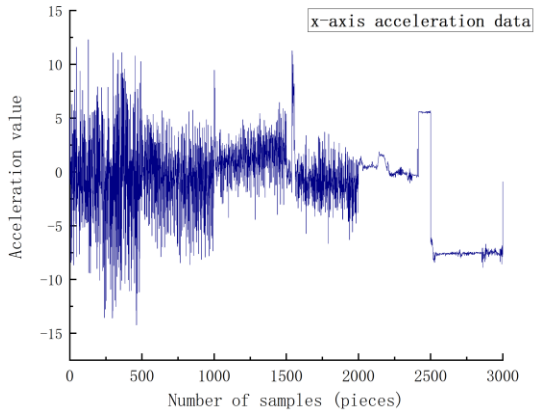

Figure 4 x-axis acceleration data

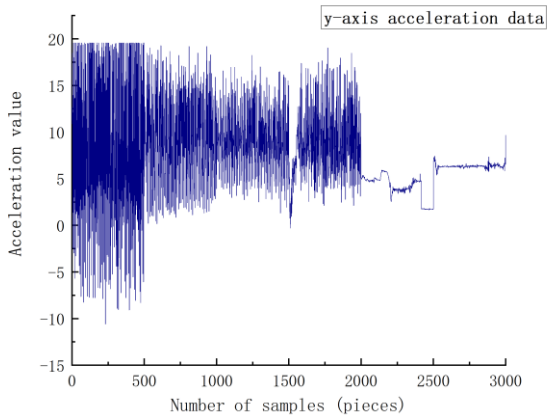

Figure 5 y-axis acceleration data

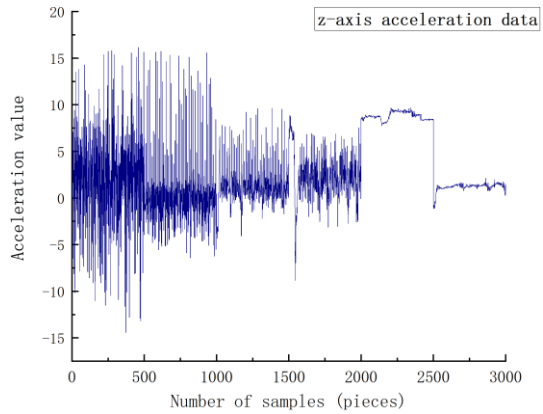

Figure 6 z-axis acceleration data

Through the observation of the above figure, we can see that there are obvious differences in the numerical fluctuation of the three-axis acceleration under different human behaviors. Since the sitting and standing movements do not move back and forth, the acceleration values are relatively gentle, and are obviously different from other movements.

3.2. Four neural network model parameters

The neural network model structure compared in this article is relatively simple, and only the network model needs to be changed under the premise that the remaining parameters are the same.

- Deep Neural Networks: a total of three fully connected layers are set up, each fully connected layer is followed by a dropout layer, the last layer is the Softmax layer, the number of nodes in each layer is 256, the dropout rate is set to 0.5, and the activation function is Relu , The number of training sessions is 100.
- Convolutional neural network: The number of convolutional layers is 1, and the convolutional layer has 16 filters of size 2*2. The number of nodes in the fully connected layer is 72, the dropout rate is 0.5, the activation function is Relu, and the number of training is 100.

- LSTM network: The number of neuron nodes in LSTM is set to 24. A fully connected layer, the Dropout rate is set to 0.5, the activation function is Relu, and the rest of the parameter settings are the same as the convolutional neural network.
- CNN-LSTM hybrid network: one layer of convolutional layer, one layer of LSTM, one layer of fully connected layer, one layer of Softmax layer, the dropout rate is set to 0.5, the activation function is Relu, and the rest of the parameter settings are the same as the convolutional neural network.

In the training of the network model, the choice of each parameter will have an impact on the results, in order to comprehensively consider factors such as accuracy, complexity, and ease of operation. We set the number of training times to 100 times, because at 100 times, the results have stabilized, and increasing the number of training times will not change the results, but will consume more computing resources; the principle of determining the number of convolutional layers is Long and deep. The higher the number of network layers, the stronger the expressive ability. However, when the number of network layers is too high, it will cause problems such as difficulty in network training and slow convergence. In order to compare the above four neural networks more effectively, the training speed will be faster while Does not affect the training results.

The number and size of the convolution kernel in the neural network are the key points. The choice of the size of the convolution kernel needs to meet two conditions: one is to be as small as possible, because a small convolution kernel requires fewer parameters and requires less computing resources; the other is that the convolution process should not change the feature map size of. Selection of the number of convolution kernels: each convolution kernel corresponds to a feature map extracted. Increasing the number of convolution kernels can increase the number of feature extractions and increase the recognition efficiency, but too many convolution kernels will bring The calculation of a large number of network parameters increases the difficulty of training.

3.3. Results and analysis

The training times of each network is 100 times. After comparing the training results of multiple networks, it can be seen that there are differences between the training results of different networks. Under the above neural network models, the test accuracy rate under the Deep Neural Networks is at least 86.7%; the test accuracy rate under the convolutional neural network is 90.1%; the test accuracy rate under the LSTM is higher than 94.8%; CNN -LSTM has the highest accuracy rate and the best effect at 96.9%.

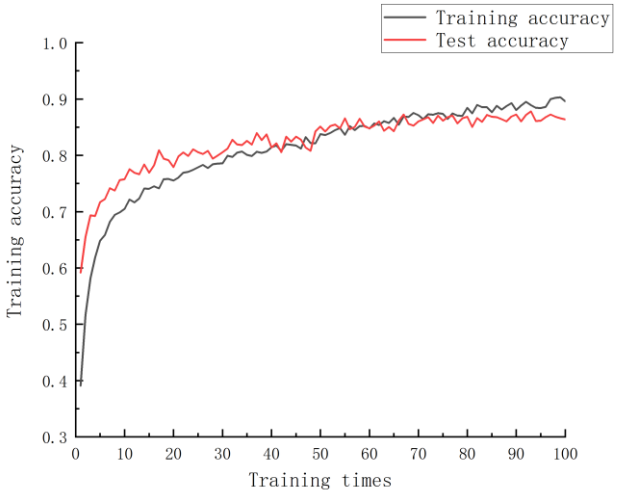

Figure 7 Test accuracy under DNN

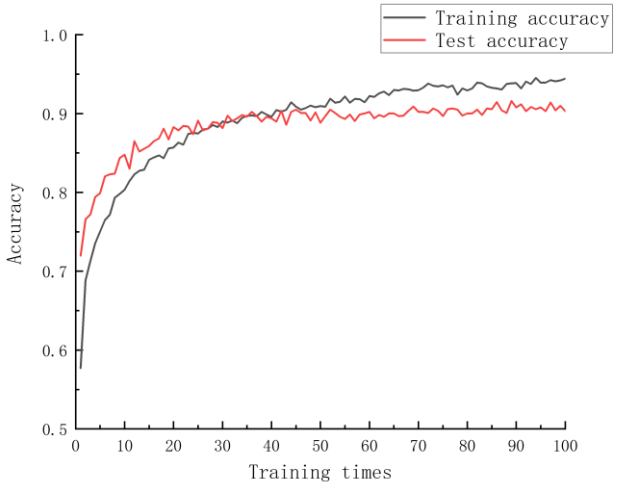

Figure 8 Test accuracy under CNN

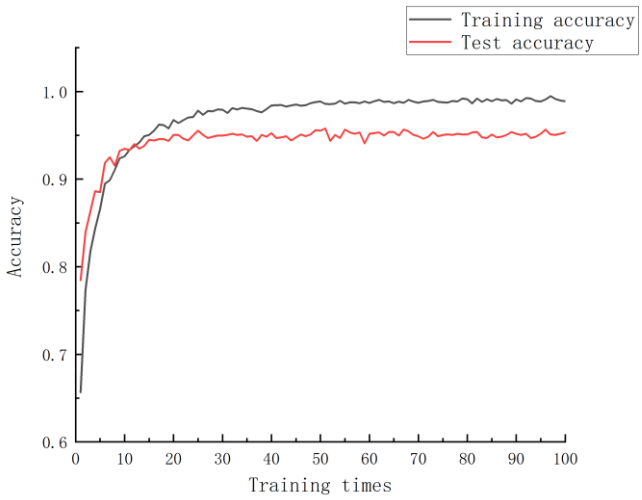

Figure 9 Test accuracy under LSTM

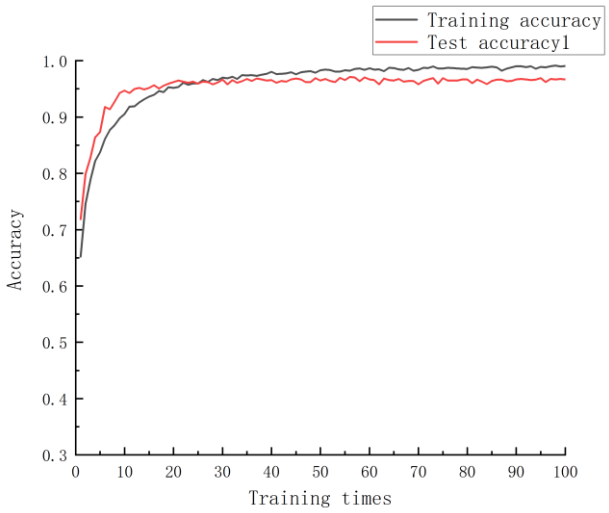

Figure 10 Test accuracy under CNN+LSTM

Table 1 Comparison of accuracy rates under four network models

Network model	Accuracy
DNN	86.7%
CNN	90.1%
LSTM	94.8%
CNN-LSTM	96.9%

The data set used in this article contains six kinds of actions, and the recognition rate of each action is also different. According to the observation of the confusion matrix in the figure below, the recognition accuracy of upstairs and downstairs in each neural network is lower than that of other actions. This is because the difference between the original data of upstairs and downstairs is small. To the extent it will produce misjudgments. Among the four kinds of neural networks, the Deep Neural Networks has the worst training effect on human behavior. This is because the Deep Neural Networks has a limit on the number of layers, and the parameters will increase exponentially after passing through the fully connected layer. The amount of calculation is large, and the learning efficiency is low. Convolutional neural network and long short-term memory network are both single network models, but in comparison, LSTM has better recognition accuracy. The motion behavior of the human body is related, and the front and back movements are continuous, and LSTM has a unique advantage in this respect and can be better classified.

The training accuracy of the CNN-LSTM hybrid model is 96.9%, and the effect is better than that of CNN and LSTM alone. The accuracy of the six actions has been improved. The recognition accuracy of jogging and standing can reach 100%. Although the accuracy of going upstairs and going downstairs only reached 92% and 93%, compared with other network models, Has reached a substantial increase.

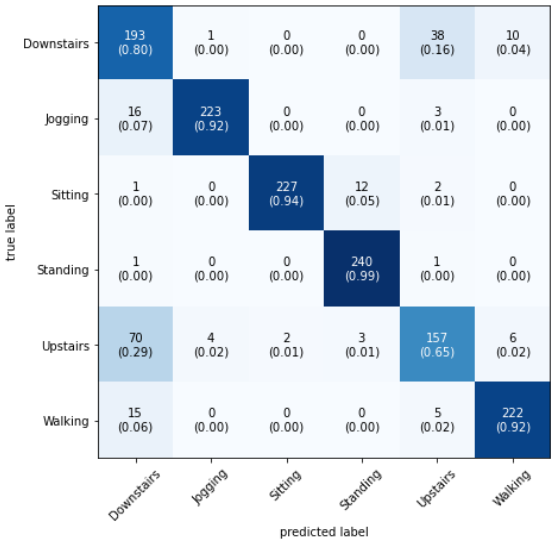

Figure 11 DNN's confusion matrix

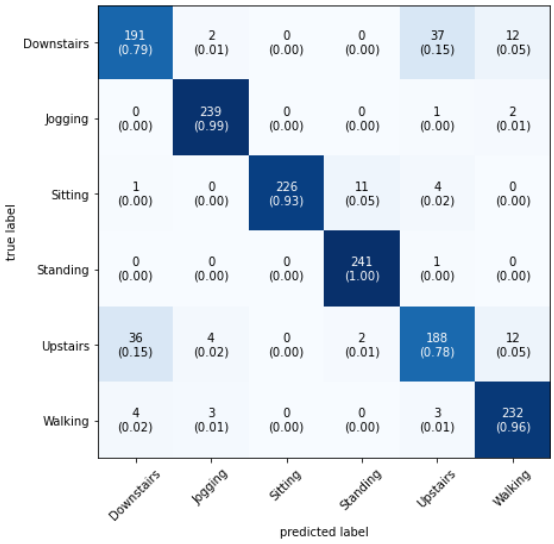

Figure 12 CNN's confusion matrix

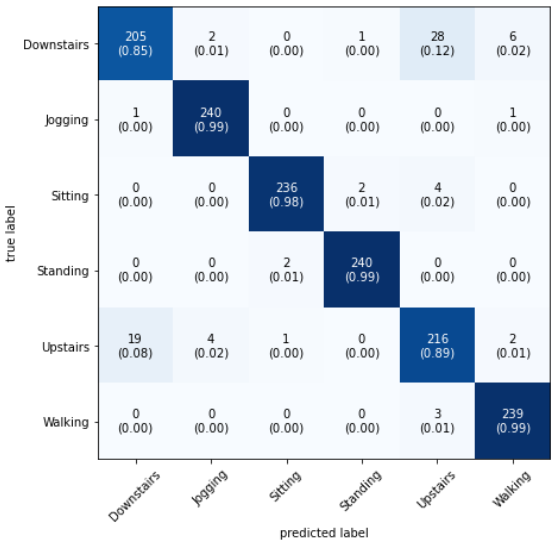

Figure 13 LSTM's confusion matrix

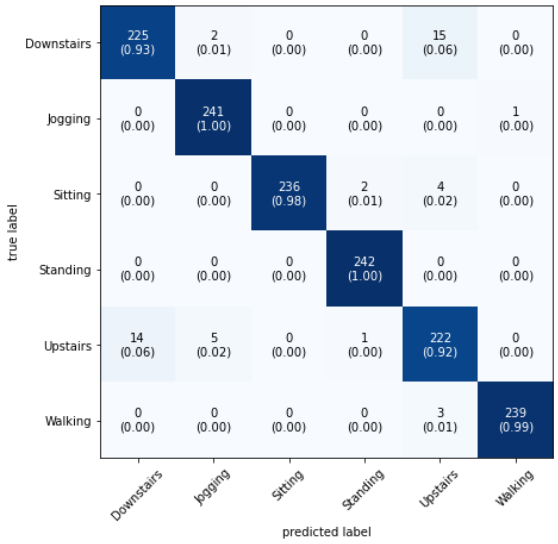

Figure 14 CNN-LSTM's confusion matrix

The following table shows the comparison of the recognition rates of the six actions in each network.

Table 2 Comparison of the recognition rate of six actions in each network

	DNN	CNN	LSTM	CNN-LSTM
Downstairs	80%	79%	85%	93%
Jogging	92%	99%	99%	100%
Sitting	94%	93%	98%	98%
Standing	99%	100%	99%	100%
Upstairs	65%	78%	89%	92%
walking	92%	96%	99%	99%

4. CONCLUSION

This paper compares the recognition results of DNN, CNN, LSTM and CNN-LSTM under accelerometer-based human behavior recognition. It can be seen from the test results and the confusion matrix that the CNN-LSTM hybrid model is better than the single neural network model. The advantage of CNN-LSTM is that it cannot only automatically extract features, but also fully consider the behaviors in the previous and later time series in the prediction stage of the network model, so that the final network prediction results no longer rely solely on features to classify, experiment The results also show that this combination method can effectively improve the recognition rate of actions, and it is a method worth continuing to explore.

ACKNOWLEDGEMENTS

This work was supported by the National Key RESEARCH and Development Program No.2020YFC1511702, the National Natural Science Foundation of China No.61801032, the Natural Science Foundation of Beijing, and the key RESEARCH Nos.4214071, 4212003 and development project of the Science and Technology Commission of the Military Commission: M Intelligent Navigation Theory research.

REFERENCES:

[1] Zhang Xingjian. Recognition and Development of Human Behavior[J]. Journal of Shanghai University of Electric Power, 2017(1).

[2] Ming Z , Le T N , Bo Y , et al. Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors[C]// 6th International Conference on Mobile Computing, Applications and Services. IEEE, 2015.

[3] Zhang L , Wu X , Luo D . Recognizing Human Activities from Raw Accelerometer Data Using Deep Neural Networks[C]// IEEE International Conference on Machine Learning & Applications. IEEE, 2015:865-870.

[4] Muscillo R , Schmid M , Conforto S , et al. An adaptive Kalman-based Bayes estimation technique to classify locomotor activities in young and elderly adults through accelerometers[J]. Medical Engineering & Physics, 2010, 32(8):849-859.

[5] Bevilacqua A, MacDonald K, Rangarej A, et al. Human activity recognition with convolutional neural networks[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2018: 541-552.

[6] Park J H, Park S U, Uddin M Z, et al. A single depth sensor based human activity recognition via convolutional neural network[C]//International Conference on the Development of Biomedical Engineering in Vietnam. Springer, Singapore, 2017: 541-545.

[7] Li Songling. Research on Human Action Recognition Based on Convolutional Neural Network[D]. University of Electronic Science and Technology of China, 2019.

[8] Wu Jun,Xiao Kecong. Human action recognition based on deep convolutional neural network[J].Journal of Huazhong University of Science and Technology (Natural Science Edition),2016,44(S1):190-194.

[9] Yang Shiwen. Research on Action Recognition Method Based on Wearable Devices [D]. Fuzhou University, 2018.

[10] Yang Heng, Yue Jianping, Xing Yin, Zhou Qinkun. Research on dam deformation prediction based on deep fully connected neural network [J]. Geodesy and Geodynamics, 2021, 41(02): 162-166.

[11] Chen Fei, Cheng Hebin, Wang Weiguang. Human behavior recognition method based on CNN-LSTMs hybrid model[J]. Information Technology and Information Technology, 2019(04): 32-34.

[12] Zou Xiaowu, Sheng Mengmeng, Mao Jiafa, Sheng Weiguo. A CNN-BLSTM model for human behavior recognition [J]. Small Microcomputer System, 2019, 40(11): 2313-2317.

[13] Han Xinxin, Ye Jian, Zhou Haiying. CNN method based on data fusion for human activity recognition [J]. Computer Engineering and Design, 2020, 41(02): 522-528.