doi:10.3772/j.issn.1006-6748.2021.04.006

Behavior recognition algorithm based on the improved R3D and LSTM network fusion①

Wu Jin (吴 进)②, An Yiyuan, Dai Wei, Zhao Bo
(School of Electronic and Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, P. R. China)

Abstract

Because behavior recognition is based on video frame sequences, this paper proposes a behavior recognition algorithm that combines 3D residual convolutional neural network (R3D) and long short-term memory (LSTM). First, the residual module is extended to three dimensions, which can extract features in the time and space domain at the same time. Second, by changing the size of the pooling layer window the integrity of the time domain features is preserved, at the same time, in order to overcome the difficulty of network training and over-fitting problems, the batch normalization (BN) layer and the dropout layer are added. After that, because the global average pooling layer (GAP) is affected by the size of the feature map, the network cannot be further deepened, so the convolution layer and maxpool layer are added to the R3D network. Finally, because LSTM has the ability to memorize information and can extract more abstract timing features, the LSTM network is introduced into the R3D network. Experimental results show that the R3D + LSTM network achieves 91% recognition rate on the UCF-101 dataset.

Key words: behavior recognition, three-dimensional residual convolutional neural network (R3D), long short-term memory (LSTM), dropout, batch normalization (BN)

0 Introduction

Due to the increasingly high status of video human behavior recognition in the field of artificial intelligence, people's demand for behavior recognition intelligent system is growing. Therefore, video based behavior recognition is widely used in human-computer interaction, social public security, intelligent security and other fields[1]. Currently, the traditional algorithms for human behavior recognition include histogram of optical flow (HOF)[2], dense trajectory (DT)[3], motion history image (MHI)[4] algorithm. Scale invariant feature transform (SIFT)[5], space-time volume (STV)[6] and dense trajectories (DT)[7] proposed by other scholars are classified after feature extraction.

In recent years, with the increase in the number of videos, the computer performance has improved rapidly, which has brought great help to the development of deep learning, and solved the problems of less data sets and slow computing performance. After Krizhevsky et al.[8] won the champion in Imagenet Challenge Image Classification, a large number of scholars began to imitate the convolutional neural networks (CNN) model, and a large number of excellent network models such as AlexNet[9], VGGNet[10], GoogLeNet[11] were proposed.

In order to enable CNN to achieve end-to-end training, Ref.[12] proposed a long-term recurrent neural network (LRCN) in 2015. This model has obvious advantages in recognition, optimization and other tasks. However, because the number of layers of CNN is too small, it can not fully extract useful feature information. Ref.[13] proposed a 3D-CNN, which can simultaneously extract spatiotemporal features, but 2D-CNN is still used in the last few layers of the network. Ref.[14] proposed a C3D network. Experimental results show that the C3D network can extract spatiotemporal feature information better than 2D-CNN. However, as the number of network layers becomes deeper, problems such as network degradation will occur. Ref.[15] proposed a ResNet network, which overcomes the above problems caused by increasing the network depth.

In order to improve the performance of the net-

① Supported by the Shaanxi Province Key Research and Development Project (No. 2021GY-280), Shaanxi Province Natural Science Basic Research Program (No. 2021JM-459), and the National Natural Science Foundation of China (No. 61772417).
② To whom correspondence should be addressed. E-mail: wujin1026@126.com
Received on Dec. 1, 2020

work, this paper introduces the three-dimensional residual convolutional neural network (R3D), which can not only extract the temporal and spatial features, but also deepen the width of the network. On this basis, R3D network changes the size of the pool layer window, and adds Softplus activation function, batch normalization (BN) layer, dropout layer, convolutional layer and maxpool layer. Later, in order to further extract advanced timing features, the long short-term memory (LSTM)[16] network was introduced into R3D network. Finally, the R3D + LSTM network achieves 91% recognition rate on UCF-101[17] dataset.

1 R3D + LSTM network

1.1 R3D network

The structure of the residual network is to emulate the VGGNet, using a small convolution kernel instead of a large convolution kernel, reducing the amount of parameters. Moreover, through residual connection, the network layers are stacked to 152 layers, which has achieved good results in Imagenet competition. The residual module is shown in Fig. 1.

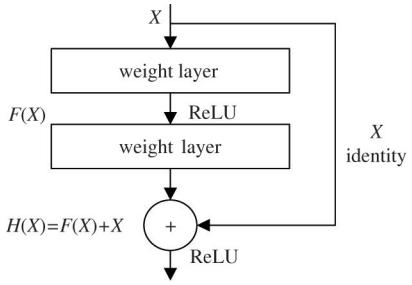


Fig. 1 Residual module

The objective function $H(X) = F(X) + X$, $F(X)$ is fitted to 0, that is, $H(X) = X$, which is transformed into the fitting of the network to X, realizing the identity mapping of X, solving the problem of network degradation. Since the derivative of X is 1, the derivative value of the function is made greater than 1 in the backpropagation, which avoids the disappearance of the network gradient and makes the weight of the network updated. Because the number of layers of the traditional deep ResNet network is too deep, there are problems such as excessive parameter amount and redundant parameters, which causes the training speed of the network to slow down[18]. Moreover, the ResNet network uses 2D convolution layer, which can only extract the spatial features of each image frame, so that the extracted features are not enough. In view of the above problems, this paper adopts R3D network, as shown in Fig. 2.

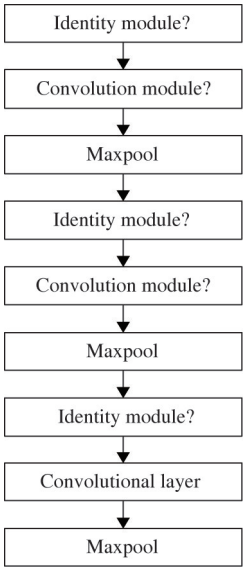


Fig. 2 Structure diagram of R3D network

Since the operations and parameters of the 5 identity modules Ⅰ are the same, one identity module Ⅰ is used to represent the 5 identity modules Ⅰ in the R3D network structure. Four identity modules Ⅱ and 4 identity modules Ⅲ are also represented in this way. The identity module uses $3 \times 3 \times 3$ convolution kernel, and the convolution module uses $3 \times 3 \times 3$ and $1 \times 1 \times 1$ convolution kernel. The above convolutional layers all use Softplus to replace the ReLU activation function, because the value of the ReLU function in the negative interval is 0, so that some neurons cannot be activated, and therefore, the corresponding weight parameters cannot be updated.

1.2 LSTM network structure

As an improved version of recurrent neural network (RNN)[19], LSTM has a very good effect on processing video, which has time-dimensional feature. It perfectly solves the problem of long-term dependence of RNN. The key of LSTM is the state of each cell, as shown in Fig. 3.

Among them, l_i is an element in the input sequence $\{l_1, l_2, l_3, \cdots, l_{n-1}, l_n\}$, and the sequence length is n. LSTM, like RNN, needs to calculate the current hidden state h_t, the hidden layer state can extract the feature of the sequence data, and then convert them to output. Use h_t to represent the hidden layer state of l_i at different times. The hidden layer state is related to the previous historical information.

In the human behavior recognition task of video class, each category of video is converted into hundreds of frames. In this paper, the number of consecutive frames input each time is a sequence of n video frames, and the output is the corresponding video category.

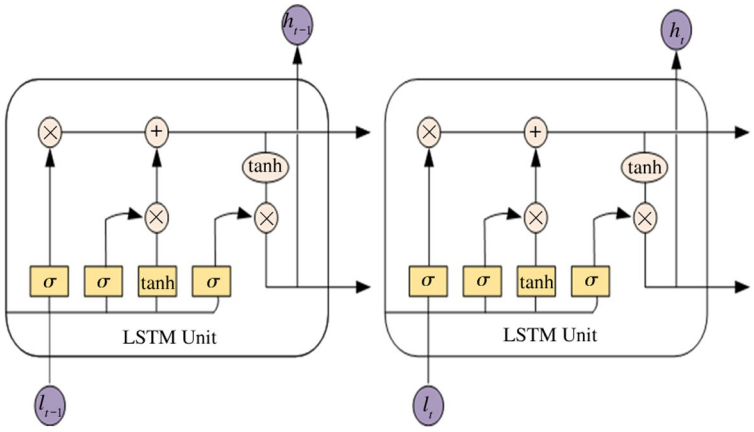


Fig. 3 LSTM neuron connection method

Therefore, the last hidden layer state h_n is selected as the high-level feature of the entire video frame. The specific calculation formula of time sequence of LSTM network is shown in Eq. (1).

$$h_n = f(Ul_n + Wh_{n-1} + b) \quad (1)$$

Among them, the bias value is represented by b, and the weight value is represented by W and U.

1.3 R3D + LSTM network structure

Ref. [8] used the CNN + LSTM method to design the network model to further improve the network classification effect, but with the increase in the number of network layers, gradient dispersion will occur, so this paper proposes the R3D + LSTM network. The network convergence architecture diagram is shown in Fig. 4.

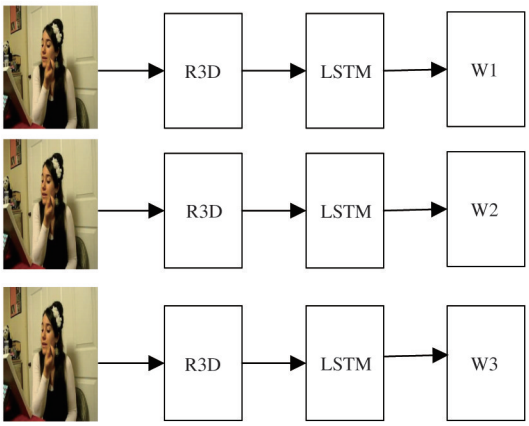


Fig. 4 Structure diagram of R3D + LSTM network

Firstly, R3D network compresses and extracts time domain features. Global average pooling (GAP) network layer further compresses model parameters to avoid over fitting of network and speed up training speed, but it can not process time domain features well. Secondly, the depth of R3D + LSTM network is not enough, which leads to a small improvement in recognition rate. Thirdly, the maxpool layer will lose a lot of useful sequence information after downsampling. Therefore, in view of these three problems, the R3D network is modified as follows.

(1) Since the GAP network is affected by the size of the feature map, the network can not be further deepened, and larger features will lead to smaller receptive field of convolution layer. Therefore, on the basis of the R3D network, convolutional layer and maxpool layer are added to deepen the depth of the network, improve the generalization ability of the network, enlarge the receptive field of convolution layer, and extract features.

(2) Rewrite all the sampling windows of the maxpool layer of the R3D network from ($2 \times 2 \times 2$) to ($1 \times 2 \times 2$) to maintain the features extracted by the shallow network and keep the time-domain sequence features intact. It avoids the loss of useful feature information when the pooling layer is down sampling.

In the dimension of input data, the feature map is expanded into one dimension, and all information features are directly input into LSTM network for feature screening, which can retain important features.

1.4 Overall network structure design

R3D + LSTM network uses identity module, convolution module, BN, Dropout and LSTM algorithm. The network has 34 convolutional layers, of which the identity module has 26 convolutional layers and the convolutional module has 6 layers. The following details the network layer structure.

Identity module I uses two 3D convolution layers to extract features, which are conv3d_2, conv3d_3, as shown in Fig. 5. Each convolution layer contains 128 convolution cores with a size of $3 \times 3 \times 3$. After that, BN layer and softplus layer are added after the convolution layer. The BN layer only normalizes the input data in batches, and the softplus function only performs nonlinear processing. Therefore, the size of the

output feature graph is $16 \times 128 \times 128$. The final output result is addition of the outputs of two convolution layers and the input of identity module I to obtain. None $\times 16 \times 128 \times 128 \times 128$, which also reflects the meaning of R3D network residual module.

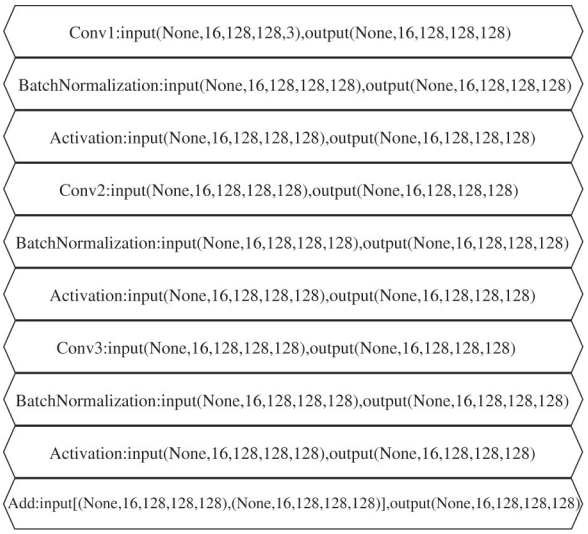


Fig. 5 Structure diagram of identity module I

The convolution module I structure contains 3 convolution layers, which are conv3d _ 12, conv3d _ 13 and conv3d _ 14. The size of the conv3d _ 12 and conv3d _ 13 convolution kernels is the same as that of the identity module I, and the number of convolution core is twice that of identification module I, so more image features can be obtained. The difference between convolution module I and identity module I is that the input data has to be processed by conv3d _ 14 convolution operation. If adding by add, the premise is that the input feature map size and the number of channels are the same. Since the stride size of conv3d _ 12 is 2, the size of the output feature map becomes 1/2 of the original size, which is $8 \times 64 \times 64$. At the same time, the stride size of conv3d _ 13 is 1, so the feature map size remains unchanged. While the conv3d _ 14 convolution kernel size is $1 \times 1 \times 1$, and the stride size is 2, which reduces the amount of parameter calculation, as shown in Fig. 6.

Each convolution module is connected with the maxpool layer to remove the lower value of the activation function response in the local neighborhood, which can reduce the dimension.

Because the GAP network is affected by the size of the characteristic graph, the network can not be further deepened, so GAP layer is removed and a layer of convolution layer and maxpool layer are added to deepen the network depth. Then, in order to further improve the network performance, LSTM network is introduced into R3D network, as shown in Fig. 7.

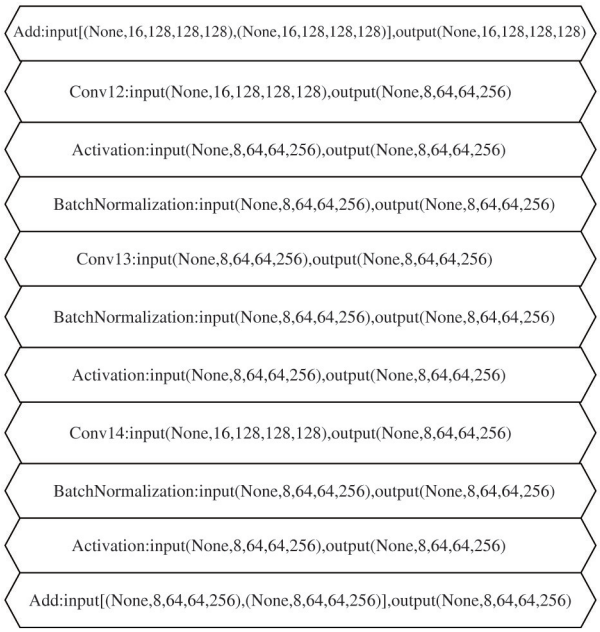


Fig. 6 Structure diagram of convolution module I

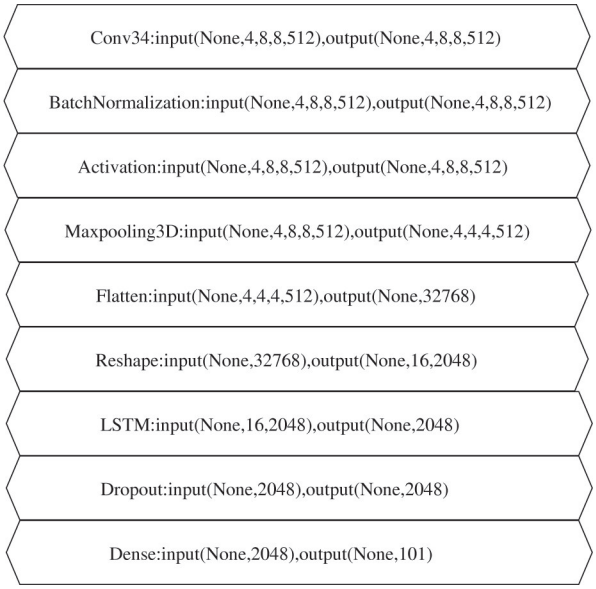


Fig. 7 Converged network structure diagram

2 Experiment and analysis

2.1 Experimental environment

The experimental environment of R3D + LSTM network is listed in Table 1.

2.2 UCF-101 dataset

The dataset used in this paper is UCF-101. This dataset contains 13 320 human behavior videos (each video is 5 – 10 s long), including 101 categories, as shown in Fig. 8.

Table 1 Experimental environment

Experimental hardware and software environment and description
1. Operating system, Ubuntu 18.04, Linux version
2. CUDA10.1, the underlying software platform for GPU acceleration
3. Keras version 2.1.4, the backend uses the TensorFlow framework
4. TensorFlow version 1.14.0
5. Graphics card, GTX1080Ti, 11 GB video memory
6. Memory, Kingston HyperX Savage DDR4, main frequency: 2400 MHz, capacity: 8 GB
7. Hard disk, Western Digital, 1000 GB memory, SATA interface, maximum transfer rate 300 MB/s.

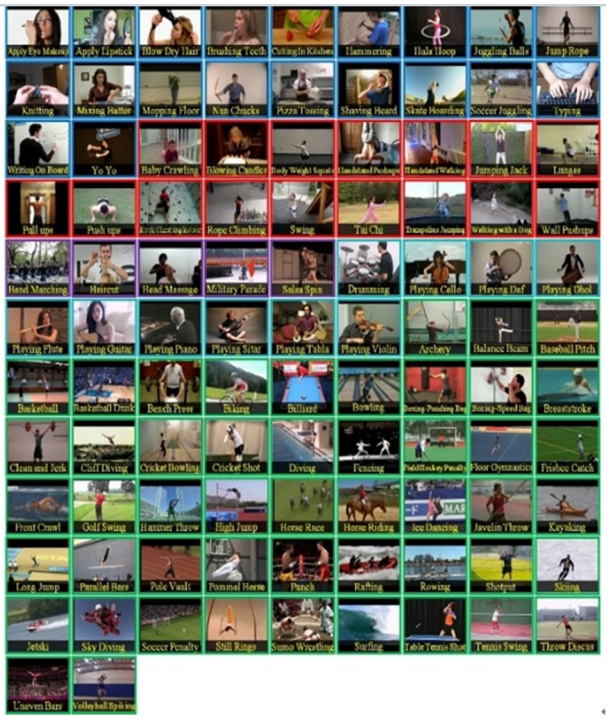


Fig. 8 All categories of UCF101 dataset

2.3 Experimental data preprocessing

Since it is not advisable to input video directly into the network, it is necessary to convert the video into a sequence of picture frames, which can speed up the training of the network. First, 13 320 videos in the UCF-101 dataset are converted, and then the naming of each converted image sequence is determined by the sequence in the video. After that, because the total number of images is too much, if all the images are input into the network at one time, the network calculation will be too large. Therefore, this paper uses the sequence with length of 16 as the input data, selects the sequence with the length of R, and then randomly generates L between $(0, R-16)$, which is used as the starting frame, and then the ending frame is selected in $(L, L+16)$. This not only prevents data from being missed, but also avoids repeating training of the same data.

2.4 Analysis of experimental results

In order to improve the training speed of the network, this paper uses an initial learning rate of 0.001. When each cycle is 24 000 times, the learning rate is reduced to 1/2 of the original, cycle 10 times, a total of 240 000 times. The hyper-parameters of the network are shown in Table 2.

Table 2 SE-R3D network hyper-parameters

Parameter	Value
base _ lr	0.01
batch _ size	16
steps _ per _ epoch	600
epochs	40
Stepsize	24 000
max _ iter	240 000
momentum	0.9
decay	1e-6

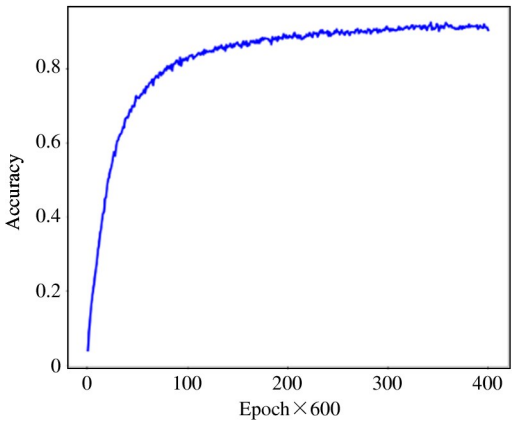
There are 400 epochs in the network, and each epoch iterates 600 times. At the 250th epoch, the convergence speed of the network begins to slow down. At the 300th epoch, the network has basically converged. At this time, the number of iterations is 180 000. Finally, R3D + LSTM network achieves 91% accuracy, as shown in Fig. 9.

The typical category accuracy rate of R3D + LSTM network on UCF-101 dataset is shown in Fig. 10. The algorithm achieves more than 90% on Cleanandjerk and Cliffdiving, more than 80% on Skydiving and Throwdiscus, while the recognition rate in the category of Blowdryhair is low, 60%. Therefore, it can be found that the accuracy of single action is usually higher than that of a complex action.

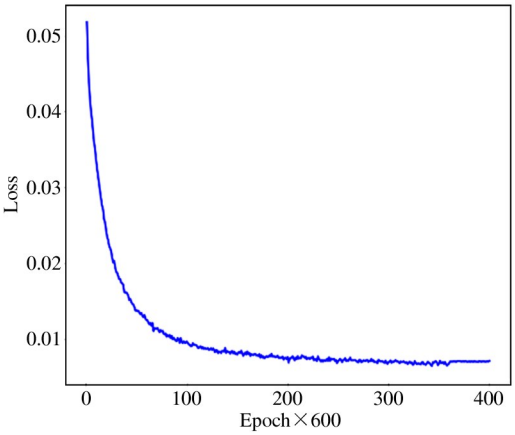
The accuracy of R3D + LSTM network is compared with other networks on UCF-101 dataset, as shown in Table 3.

Table 3 Comparison of accuracy

Network	UCF-101 dataset
C3D	85.2%
C3D + IDT	90.2%
LRCN	82.92%
Two-Stream-I3D	98%
DMC + Net[20]	90%
R3D + LSTM	91%



(a) Accuracy curve during



(b) Loss function curve

Fig. 9 R3D + LSTM training process

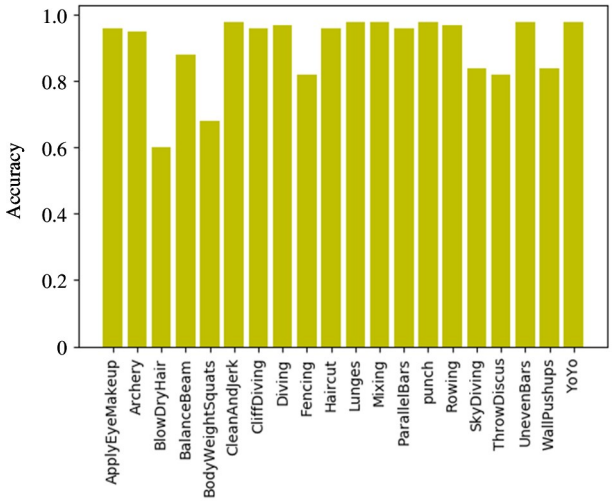


Fig. 10 Test accuracy curve

It can be seen from Table 3 that the Two-Stream-I3D model has achieved a 98% recognition rate on the UCF-101 dataset. Although the accuracy of the network designed in this paper is not as high as Two-Stream-I3D. However, compared with the popular C3D and C3D + IDT networks in the past two years, R3D + LSTM has a greater improvement in the recognition rate, and at the same time, the recognition rate is 1% higher than that of the DMC-Net network. Secondly, the recognition rate of R3D + LSTM network is much better than that of the LRCN network, which shows that the combination of the three-dimensional residual network and the LSTM network is feasible in the field of behavior recognition.

3 Conclusions

Automatic recognition of behavior in video is a long-term goal of computer vision and artificial intelligence. In order to improve the network performance, this paper designs R3D + LSTM network. First, the R3D network is modified, the ReLU activation function with Softplus is replaced, and a convolutional layer and maxpool layer is added to increase the depth of the network. Then, the pooling window of all maxpool layers is changed to (1, 2, 2) to maintain the features extracted by the shallow network, and BN layer and Dropout layer are added to improve the convergence speed of the network and effectively restrain over fitting. Later, in order to extract the high-level temporal features, LSTM network is introduced. Finally, the R3D + LSTM network achieves 91% recognition rate on the UCF-101 dataset.

Although the R3D + LSTM network designed in this paper has achieved good performance in recognition rate, compared with some perfect algorithms in this field, there is still room for improvement. The future work and prospects are as follows.

(1) Optimization of the model. The designed network model can be further optimized to obtain a higher recognition rate, and more datasets will be used to test the performance of the model.

(2) The datasets used are preprocessed, but in actual scene, the behavior will become more complex and the resolution of the video will be reduced. Therefore, further research needs to be done to identify the human behavior categories accurately and efficiently.

Reference

[1] Wu J, Min Y, Shi Q W, et al. Behavior recognition based on the fusion of 3D-BN-VGG and LSTM network [J]. *High Technology Letters*, 2020, 26(4):372-382

[2] Brox T, Malik J. Large displacement optical flow: descriptor matching in variational motion estimation [J]. *IEEE Transactions on Pattern Analysis and Marchine Intelligence*, 2011,33(3): 500-513

[3] Wang H, Klser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition [J]. *International Journal of Computer Vision*, 2013,

103(1):60-79
[4] Li D X, Yu L J, He J, et al. Action recognition based on multiple key motion history images[C]// IEEE International Conference on Signal Processing, Chengdu, China, 2016:993-996
[5] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004,60(2):91-110
[6] Laptev I. On space-time interest points[J]. *International Journal of Computer Vision*, 2005, 64(2-3):107-123
[7] Wang H, Schmid C. Action recognition with improved trajectories[C]// Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 2013:3551-3558
[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012:1097-1105
[9] Zamri N N M, Ling G F, Han P Y, et al. Vision-based human action recognition on pre-trained AlexNet[C]// 2019 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 2019:1-5
[10] Yang Z Q. Gesture recognition based on improved VGG-NET convolutional neural network[C]//2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 2020:1736-1739
[11] Aswathy P, Siddhartha, Mishra D. Deep GoogLeNet features for visual object tracking[C]//2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, 2018:60-66
[12] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 4(39): 677-691
[13] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(1): 221-231
[14] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4489-4497
[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Computer Vision & Pattern Recognition, Washington DC, USA, 2016:770-778
[16] Shu X, Zhang L, Sun Y, et al. Host-parasite: graph LSTM-in-LSTM for group activity recognition[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 5(99):1-12
[17] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human action classes from videos in the wild[J]. *arXiv*:1212.0402, 2012
[18] Zheng Y, Wang R G, Yang J, et al. Principal characteristic networks for few-shot learning[J]. *Journal of Visual Communication and Image Representation*, 2019, 59(3): 563-573
[19] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. *Nature*, 1986,323(6088): 533-536
[20] Shou Z, Zhi C Y, Yannis K, et al. DMC-Net: generating discriminative motion cues for fast compressed video action recognition[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019:1268-1277

Wu Jin, born in 1975. She received her B.S. and M.S. degrees from Xi'an Jiaotong University in 1998 and 2001 respectively. She has been a professor at School of Electronic Engineering, Xi'an University of Posts and Telecommunications since 2014. Her research interests include the signal and information processing.