IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 26, 2021, accepted October 1, 2021, date of publication October 13, 2021, date of current version October 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3119609

Behavior Recognition Algorithm Based on the Fusion of SE-R3D and LSTM Network

JIN WU, YI YUAN AN, QIAN WEN SHI, AND WEI DAI
School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Corresponding author: Jin Wu (wujin1026@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61834005, Grant 61772417, Grant 61802304, Grant 61602377, and Grant 61634004; and in part by the Shaanxi Province International Science and Technology Cooperation Project under Grant xyt2018KW-006.

ABSTRACT In view of the fact that the existing behavior recognition algorithms cannot fully extract abstract behavior features, this paper proposes a SE-R3D-LSTM behavior recognition algorithm based on 3D residual convolutional neural network (R3D), which integrates Squeeze-and-excitation network (SENet)and long short-term memory (LSTM). First of all, a residual module is added to the 3D Convolutional Neural Network (3D-CNN) to avoid problems such as gradient dispersion caused by the deepening of the network layer; Secondly, not only the global average pooling layer but also the global maximum pooling layer is used in the SENet network, which can fully extract global information and achieve feature calibration. In the meantime, expand the SENet network to three-dimensional, which can make the connection of the spatiotemporal feature channels closer. Afterwards, the 3D-SE module is introduced into the R3D network, which can enhance the effective spatiotemporal features and suppress the invalid spatiotemporal features; Since, because LSTM can perform timing modeling on high-level features and learn more effective feature information, the LSTM network is introduced into the SE-R3D network. Finally, Softmax is used for classification. Experimental results show that the recognition rate of the SE-R3D-LSTM network on the UCF101 data set reaches 96.5%.

INDEX TERMS Human behavior recognition, 3D residual convolutional neural network, SENet network, long short-term memory network.

I. INTRODUCTION

In the 1970s, Professor Johansson [1] proposed a description method of the human body model structure. After that, the discriminant models based on human behavior are improved on this basis [2]. Currently, the traditional algorithms for human behavior recognition include Histogram of Oriented Gradients (HOG) [3], Histogram of Optical Flow (HOF) [4], Dense Trajectory (DT) [5], and Motion History Image (MHI) [6]. In addition, cale Invariant Feature Transform (SIFT) [7], Space-time Volume (STV) [8], Local Binary Patterns (LBP) [9] and Dense Trajectories(DT) [10] etc. are also proposed by other scholoars, which are all classified after feature extraction. There are roughly two directions for the algorithm of behavior classification research, one is direct classification, such as the K-nearest neighbor algorithm [11], Support Vector Machine (SVM) [12]. The other is the time-domain state-space fusion model, such as Dynamic Time Warping (DTW) [13] and so on.

In recent years, deep learning algorithms for big data are significantly better than traditional algorithms, such as human-computer interaction, social public safety, intelligent security and other fields [14].Yann Lecun proposed convolutional neural network (CNN) [15]. Since the CNN network can only extract information from spatial domain features, but ignores temporal features. In 2014, Simonyan K et al [16] proposed a two-stream network composed of two dimensions: time and space. In the time domain network part, the optical flow field between consecutive multiple frames is used as multiple input channels, and the spatial domain network uses RGB images for training. However, the two-stream network only uses stacked video frames as multiple input channels, and does not process the video frame sequence in time sequence, so it is difficult to extract spatiotemporal motion information. In this regard, in 2015 Donahue J *et al.* proposed a Long-term Recurrent Convolutional Network (LRCN) [17], which uses CNN for static feature information extraction, and

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

the extracted features are processed by the LSTM network, However, due to the small number of layers of CNN network, useful feature information was not fully obtained in the early stage of image processing. Therefore, Ji et al [18] proposed a 3D-CNN network, which can simultaneously extract spatio-temporal features. When 2D-CNN is extended to 3D-CNN, the original two-dimensional features still represent the spatial features of static images, while the remaining three dimensions extract temporal features, and can perform convolution across multiple frames and extract the information before and after the time sequence. However, the network does not all use 3D-CNN. As the number of network layers deepens, 2D-CNN is used in the last few layers of the network. In 2015, Tran D et al [19] made improvements on the basis of the 3D-CNN network and designed a C3D network model. The convolutional layer of the C3D network model all uses 3D-CNN. Experiments show that this network is more suitable for learning space-time features than 2D-CNN. However, due to the increase in the number of parameters brings great difficulties to network training, the number of network layers cannot continue to deepen, and the increase in the amount of calculation leads to overfitting during the training process [20]. Literature [21] proposed interleaving perception convolutional neural network (IP-CNN). IP-CNN designed a two-way autoencoder to reconstruct hyperspectral and LiDAR data without involving annotation information, and then used two CNN networks to classify the fusion data, and finally achieved good results on a small-scale data set. In order to further study the classification of hyperspectral images and lidar data, the literature [22] used hierarchical random walk network (HRWN), and designed a dual-tunnel convolutional neural network based on HRWN to extract spectral and two-dimensional features, It is proved by experiments that the HRWN algorithm is better than other most advanced algorithms.Literature [23] proposed a patch-to-patch convolutional neural network (PToP CNN) and used unsupervised features to extract a framework, with the purpose of classifying hyperspectral and lidar data. Experimental verification shows that PToP CNN exhibits higher performance than two-branch CNN and context CNN. Literature [24] proposed two branched convolutional neural networks (CNN) to process hyperspectral images (HSI) and data from multiple sensors. First, the CNN network is used to extract the two-dimensional spatial features of HIS, and then the two-dimensional spatial domain features and spectral features are fused. The experimental results show that the two-branch CNN network is more suitable for data classification.

In deep learning, it is very difficult to train a network with good performance. The difficulties comes from the performance of the machine, the size of the dataset, the depth and width of the network, etc. Based on this, He Yuming et al [25] proposed Residual Neural Network (ResNet), whose structure is to emulate the VGG-Net, but introduces the residual module between the layers to avoid the vanishing gradient and network degradation caused by the deepening of the network [26], [27]. Practice has proved that this can indeed greatly improve the performance of the network. Hu Jie *et al.* Started from the characteristic channel and proposed the SENet network, which increased the characterization ability of the network. The SENet network won the champion of the Image Classification task in the last ImageNet competition [28].

In order to further improve and optimize network performance, this paper proposes the SE-R3D-LSTM network. First of all, because the number of layers of the traditional 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 [29], [30]. Not only that, the ResNet network uses a 2D convolutional layer, which can only extract the spatial features of each frame of the image, which makes the extracted features insufficient. Therefore, combining ResNet and 3D-CNN networks to form an R3D network can better extract the feature of the space-time domain. After that, GMP was added to the SENet network, because GAP represents global information by extracting the average value of global features, while GMP represents global information by extracting the global maximum feature, so after GAP and GMP operations, it can fully extract the space-time domain global information. On this basis, SENet is extended to three dimensions and integrated into the R3D network. The purpose is to gain attention weight through learning, enhance useful features, weaken useless features, and improve the representational ability of SE-R3D network. Then, the LSTM [31] network is introduced to process the abstract timing features that the SE-R3D network cannot extract. In the end, the SE-R3D-LSTM network achieved a 96.5% recognition rate on the UCF-101 [32] dataset.

II. SE-R3D-LSTM NETWORK STRUCTURE AND ALGORITHM DESIGN

A. BUILD R3D NETWORK WITH RESIDUAL MODULE

Reference [25] shows that through many experiments, it is concluded that the number of network layers will affect the recognition effect. The accuracy of the COCO target detection dataset is increased by 28% due to the use of a sufficiently deep network. Due to the introduction of the residual module, it overcomes the problem that the accuracy rate drops caused by the excessively deep network layers.

The residual module is shown in Fig.1.

For the network degradation problem caused by deeper network layers, therefore, the objective function $H(X) = F(X) + X$ is made to fit $F(X)$to 0, that is, $H(X) = X$, thus transforming to the fitting of X and realizing the identity mapping of X. Since the derivative of X is 1, the derivative value of the function is greater than 1 in the back propagation, thus solving the problem of the network generating gradient disappearance.R3D network is adopted in this paper, as shown in Figure 2.

Since the operations and parameters of the 5 identity modules I are the same, an identity module I is used in the R3D network structure to represent these 5 identity modules I,

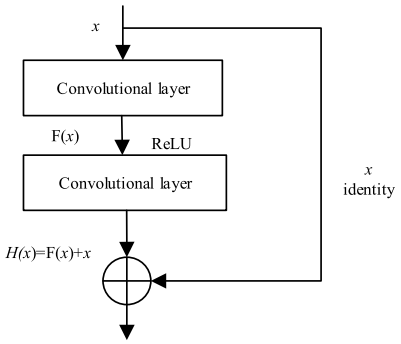


FIGURE 1. Residual module.

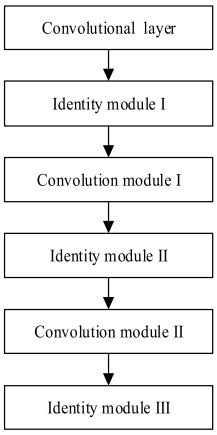


FIGURE 2. Structure diagram of R3D network.

4 identity modules II and 4 identity modules III are also this representation method. The identity modules all use a $3 \times 3 \times 3$ convolution kernel with a stride of 1, as shown in Fig.3. The convolution module uses $3 \times 3 \times 3$ and $1 \times 1 \times 1$ convolution kernels with strides of 1 and 2 respectively, as shown in Fig.4.

It can be seen from identity module I that in the training phase, the input sequence is $16 \times 32 \times 32$, 32×32 is the pixel value of the video frame, and 16 is the length of the input video sequence. The reason is that if the video sequence is too long and the video frame resolution is very high, due to the performance of the computer, the video memory will be destroyed and the training can not be carried out, or the training can be carried out, the calculation speed will be very slow. Moreover, if the video sequence is too short and the resolution is too low, it will cause too much loss of useful information, resulting in low recognition rate.The identity module I performs feature extraction through two 3D convolutional layers, which are conv3d_ 2,conv3d_ 3.BN [33] layer and ReLU layer are added after each convolution layer, And 128 convolution kernels with the size of $3 \times 3 \times 3$ are included. Because the stride of convolution layer and the edge filling of image are both 1, BN layer only does batch normalization for input data, and ReLU function only performs nonlinear processing. Therefore, the size of the output feature map is $16 \times 32 \times 32$. The final output is the sum of the output of two convolution layers and the input of identity module I,

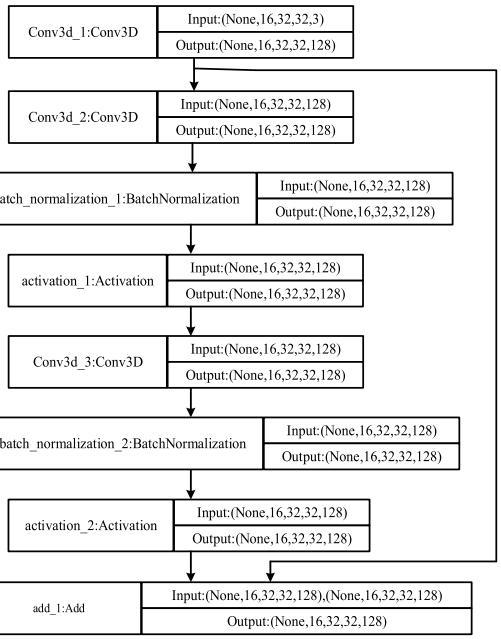


FIGURE 3. Structure diagram of identity module I.

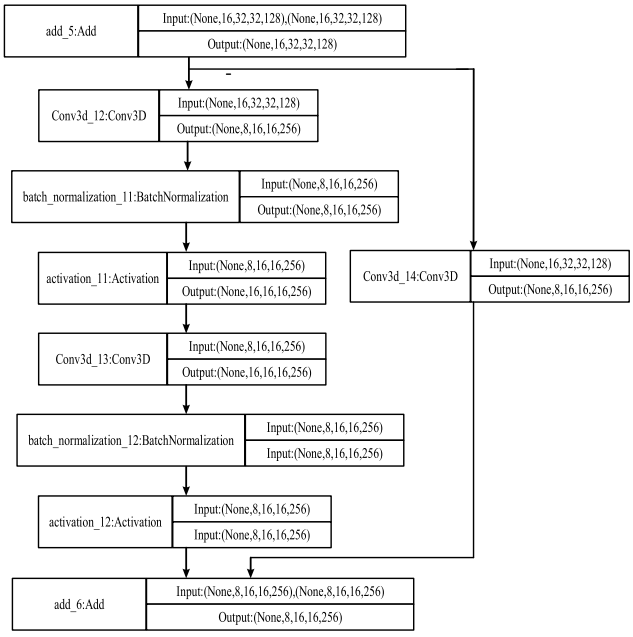


FIGURE 4. Structure of convolution module I.

that is the five-dimensional tensor of None $\times$ 16 $\times$ 32 $\times$ 32 $\times$ 128, It also reflects the meaning of R3D network residual module.

The convolution module I structure contains 3 convolution layers, which are conv3d_12, conv3d_13 and conv3d_14. The conv3d_12 and conv3d_13 convolution kernels have the same size as the identity module I, and the number of convolution kernels is twice the above, which is 256, so more image features can be obtained. And conv3d_14 The size of the convolution kernel is $1 \times 1 \times 1$, because the convolution

kernel of 1 × 1 × 1 has dual purposes. On the one hand, it increases the depth and width of the network without affecting the performance. On the other hand, it can reduce the dimension to avoid increasing the number of parameters and eliminate the calculation bottleneck. Moreover, it can choose the appropriate convolution size independently, so it can fuse the features of different scales.Because the difference between convolution module I and identity module I is that the input data has to be processed by conv3d_14 convolution operation. When adding by add, the premise is that the input feature map size and the number of channels are the same, so this paper reduces the feature map dimension by changing the stride size. Since the stride size of conv3d_12 is 2, the size of the output feature map becomes 1/2 of the original, which is 8 × 16 × 16. 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 × 1 × 1, and the stride size is 2, Which not only reduces the amount of parameter calculation but also makes the feature map size the same as the output of conv3d_13.

After that, there are 4 identity module II and identity module III, and 1 convolution module II. The number of convolution kernels is 256, 512 and 512 respectively. In addition to the difference in the number of convolution kernels, the parameters and operations of the identity module are the same, so is convolution module.

B. 3D-SENET NETWORK

1) DESIGN OF 3D-SENET NETWORK STRUCTURE

In order to increase the attention mechanism of characteristic channel in the temporal and spatial domain, this paper expands the SENet network into three dimensions. After 3D-Squeeze operation, global information can be extracted to avoid network overfitting and speed up training. Specifically, it compresses along the spatial dimension, and rotates the feature channels of the three dimensions into real numbers. In order to more fully extract the global information, the paper adds a GMP layer, because this layer can extract the global maximum feature value, it also includes the most important global information, as shown in formula (1).

$$s_c = \sigma\left(F_{ex}\left(AvgPool\left(u_c\right), W\right) + F_{ex}\left(MaxPool\left(u_c\right), W\right)\right), \tag{1}$$

where σ is the sigmoid activation function, *AvgPool* is the GAP layer and *MaxPool* is the GMP layer. The F_{ex} contains two fully connected layers. The first fully connected layer is used to reduce the dimension of input features, and the second full connection layer is used to increase the dimension, so as to better fit the complex correlation between channels.

The input of the 3D-SENet network is C × D × H × W after residual operation, C is the number of channels, D is the length of the video sequence, W and H are the width and height of the feature map. First, perform GAP and GMP operations on the input feature maps to obtain two representative global information, and the output features

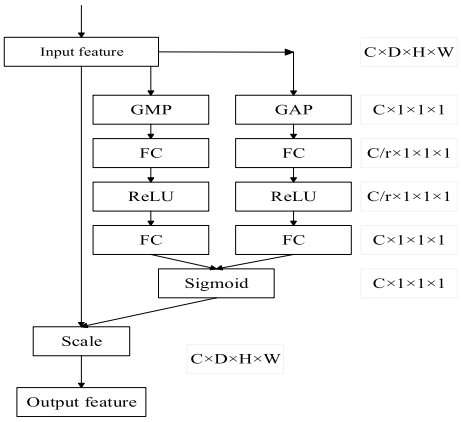


FIGURE 5. Structure diagram of 3D-SENet network.

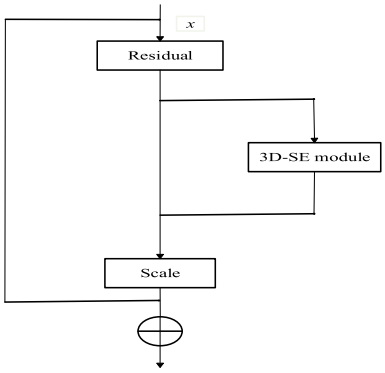


FIGURE 6. Residual unit of 3D-SE module.

are both C × 1 × 1 × 1. Then, go through the first fully connected layer to reduce the size of C to get C/r × 1 × 1 × 1. Afterwards, through the ReLU activation operation, and then through the second fully connected layer to upgrade the dimension, the output feature is C × 1 × 1 × 1. After that, the two channels are added and passed through the sigmoid function. Finally, the final output is obtained through the Scale operation, as shown in Fig.5.

It can be seen from Figure 5 that the 3D-SENet network has been feature recalibrated through Squeeze, Excitation and Scale operations, and the attention weight is obtained through learning, which improves useful spatio-temporal features and suppresses useless spatio-temporal features.

C. SE-R3D NETWORK STRUCTURE

Video-based human behavior is a series of continuous actions. If only two-dimensional convolution operations is used, only spatial features can be extracted, and the motion information between video frames in temporal dimension is ignored. The 3D-CNN can well capture the feature information of the temporal and spatial domains in the video. Therefore, the 3D convolution kernel can not only extract the spatial domain features of each frame in the video, but also extract the temporal domain features between adjacent video frames. Therefore, this paper extends the ResNet and

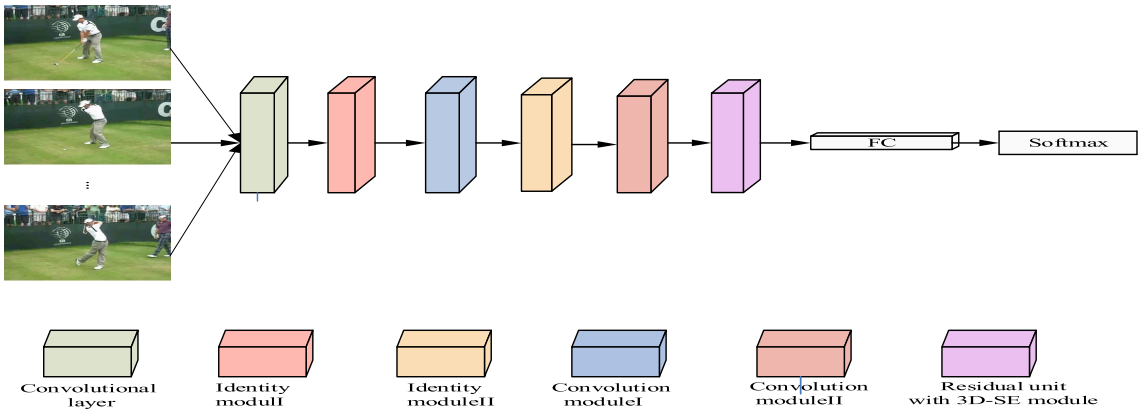


FIGURE 7. Residual unit of 3D-SE module.

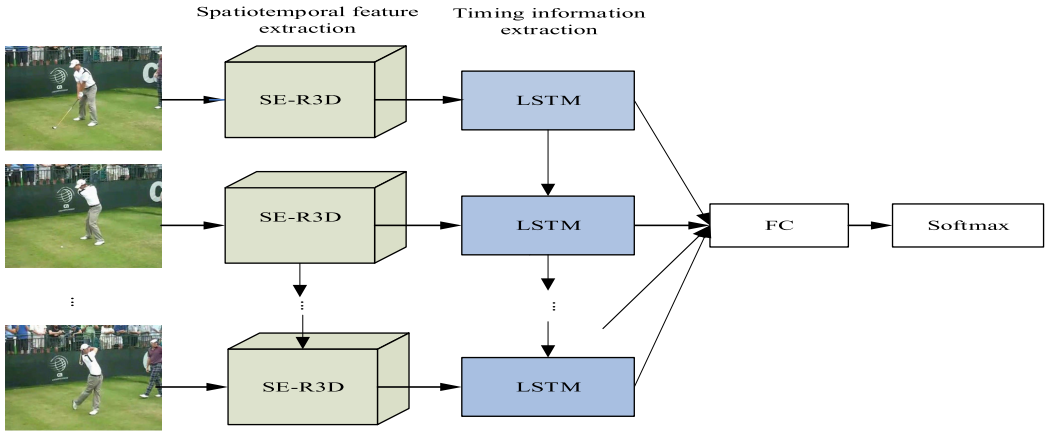


FIGURE 8. Structure diagram of SE-R3D-LSTM network.

SENet networks to three dimensions. On this basis, in order to enhance the representation ability of the network, the 3D-SE module is introduced into the residual unit to further improve the performance of the network, as shown in Fig.6.

The R3D network constructed in this paper has 6 modules from the bottom to the top. Because the shallow network has less feature information, the 3D-SE module is introduced in the last identity module III for feature calibration. Firstly, inputting the behavior recognition image sequence to the SE-R3D network for feature extraction. Then, in order to further extract effective features, FC is used to compress the feature vector, reducing the amount of parameters. Finally, using the Softmax function to classify, and get the probability of each behavior in the input image sequence, as shown in Fig.7.

D. SE-R3D-LSTM NETWORK OVERALL STRUCTURE

In order to make the recognition rate higher and the recognized feature is more accurate, the LSTM is introduced into the SE-R3D network, and then the LSTM network is used for timing modeling to learn the high-level temporal features of the video signal, as shown in Fig.8.

The R3D network only compresses and extracts the features of the time domain once. GAP layer is to further compress the model parameters, avoid overfitting of the network and speed up the training. However, it does not change much for the processing of time domain features. Secondly, the pooling layer will lose a lot of useful sequence information after down sampling. In view of these two problems, the R3D network is modified: firstly, because GAP network is affected by the size of the feature map, and the larger the feature map is, the smaller the receptive field of convolution layer is. Therefore, five identity modules I, four identity modules II and identity modules III are used in this paper, A convolution module I and convolution module II, a total of 34 layers of R3D network, to deepen the depth of the network, so that the network can have better generalization ability, at the same time, reduce the size of the feature map, making the convolution layer receptive field larger, and better feature extraction; secondly, by changing the size of the step size, the feature map is reduced to maintain the features extracted by the shallow network and make the time domain sequence special The symptoms remain intact. Secondly, by changing the stride to reduce the dimensionality of the feature map 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 performs downsampling. In order to further

TABLE 1. Experimental environment.

	The hardware and software environment of the experiment and its explanation
1:	Operating system:Ubuntu16.04.
2:	CUDA10.1:Underlying software platform for GPU acceleration.
3:	Keras edition:2.2.4, TensorFlow edition:1.14.0.
4:	Video card:GTX1080Ti,11GB Video memory.
5:	Memory:Samsung DDR4,Main frequency:2400MHZ Capacity:8GB.
6:	Hard disk: Inter SSD 545s, 256G.

improve the performance of the model, SENet is introduced, and the SE-R3D network is designed to enhance useful channel features and weaken useless channel features, and extract attention to the frame granularity of feature channels.

In the dimension of the input data, the feature map is directly expanded into one dimension,and the video frame sequence is directly connected with the LSTM, without using the full connection layer as the output of the convolution layer, and then input into the LSTM network, which can reduce the parameters of the full connection layer, speed up the calculation and prevent the network from over fitting. Secondly, all the information features are directly input into the LSTM network for feature screening, which can retain more useful features and reduce the loss of useful features.

III. EXPERIMENTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT

The experimental environment of SE-R3D-LSTM network is shown in Table 1.

B. UCF-101 DATASET

The dataset used in this paper is UCF-101, which contains 101 types of actions and 13320 videos of human behaviors, as shown in Fig.9. These videos are real scenes in real life, and each video is very different, especially the background. When extracting moving objects, the human behavior in the real world scene has nothing to do with most of the background. Therefore, extracting the moving subjects in the video sequence and using the behavior recognition network for classification can reduce the interference of background information on human behavior classification.

C. 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 rate of the network. Firstly,, 13320 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. Fig. 10 is a partial image sequence of UCF-101.

The huge data sample size is the basis for training CNN network, which can improve the generalization ability and robustness of the network model. On the contrary, the lack of sample size will lead to the problem of over fitting of

FIGURE 9. UCF-101 dataset.

FIGURE 10. Partial image sequence of UCF-101.

the network. At present, the number of samples of behavior recognition data set is still lack, and there will be a huge amount of parameters when training the network. Therefore, it is very necessary to expand the data set, but it is unscientific and time-consuming if it is necessary to collect new data artificially. At present, there are two data enhancement methods. The first is online enhancement, including data translation, flip and zoom. The second is offline enhancement, which processes data sets, but this is only for small data sets. In order to completely preserve the spatio-temporal feature information of human behavior, this paper chooses to flip the data to ensure the image quality of the extended part of the data set.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),

TABLE 2. Hyper parameters.

parameter	Value-1	value-2
base_lr	0.01	0.01
batch_size	16	16
Steps_per_epoch	600	600
epochs	20	30
Stepsize	12000	18000
max_iter	120000	216000
momentum	0.9	0.9
decay	1e-6	1e-6

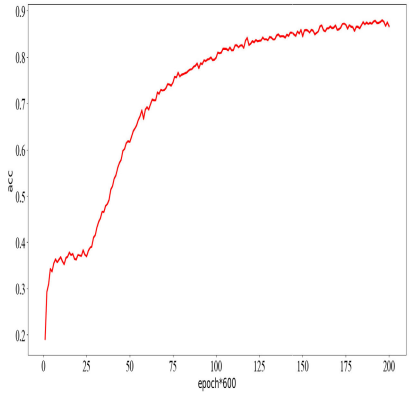
This not only prevents data from being missed, but also avoids repeated training of the same data.

D. ANALYSIS OF EXPERIMENTAL RESULT

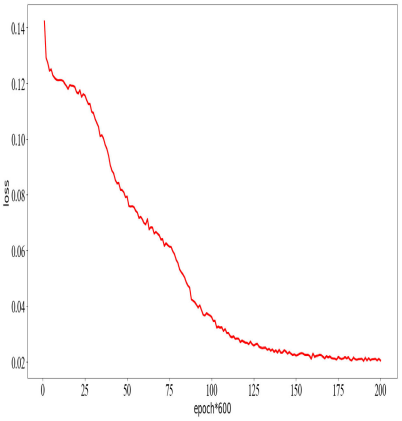
This paper uses the Ubuntu16.04 operating system and selects the SE-R3D network and SE-R3D-LSTM network built by Keras with Tensorflow as the backend. In order to improve the training speed of the network, the SE-R3D and SE-R3D-LSTM networks both use an initial learning rate of. 0.01. When each cycle is 12000 and 18000 times, the learning rate is reduced to 1/10 of the original, and the total number of cycles is 10 and 12, and the training is 120000 and 216,000 times respectively. When recording the training datas, the History module of the callback function Callbacks in Keras is used, and it is recorded once at the end of each Steps_per_epoch, and 200 and 360 data are recorded. The hyper-parameters of the two networks are shown in Table 2, where value-1 is the hyperparameter value of SE-R3D and value-2 is the parameter value of SE-R3D-LSTM.

When the SE-R3D network is trained for 3000 times, the accuracy rate curve rises slowly, and the loss rate curve decreases slowly. This is due to the high learning rate, so no local optimal solution is found. When the training reaches18000 times, the learning rate has been reduced to 0.001, which is gradually approaching the optimal solution. Therefore, the accuracy rate begins to rise significantly, the relative loss rate began to decrease significantly. When training to 90000 times, the network has converged and the recognition rate reaches 87% on UCF-101 dataset. The curve of recognition rate and loss rate of SE-R3D network model on UCF-101 dataset is shown in Fig.11. Fig. (a) is the accuracy curve, and Fig. (b) is the loss function curve.

When the SE-R3D-LSTM network is trained to 120000 times, the acc and loss values change slowly, and the network begins to converge. When the training time reaches 180000 times, the accuracy and loss rate curves begin to stabilize, and the recognition rate reaches 95% on UCF-101 dataset. The curve of recognition rate and loss rate of SE-R3D-LSTM network model on UCF-101 data set is shown in Fig.12. Fig. (a) is the accuracy curve, and Fig. (b) is the loss function curve.



(a) Accuracy curve

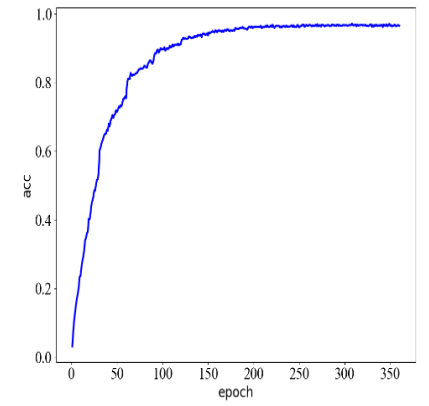


(b) Loss function curve

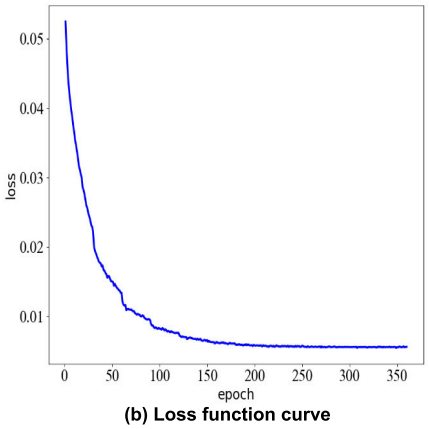
FIGURE 11. SE-R3D training process.

It can be seen from Fig. 11 and Fig. 12 that SE-R3D-LSTM has a higher recognition rate than SE-R3D network. Specifically, this paper use a 3D-CNN network model to enhance the network's ability to learn low-level spatiotemporal features. Later, as the number of network layers continues to increase, the network's representational ability will become stronger, but problems such as network degradation will occur. Therefore, this paper combines R3D and 3D-CNN networks to greatly improve the performance of the network. Then, GMP was introduced into the SENet network, because GMP also contains important global information. At the same time, the SENet network was extended to three dimensions and introduced into the R3D network, and a residual unit with a 3D-SE module was obtained, which realized the feature recalibration. Subsequently, because the LSTM network can improve the SE-R3D network's ability to extract video sequence features, the paper combines the LSTM with the SE-R3D network to increase the accuracy of the network. Finally, the SE-R3D and SE-R3D-LSTM networks have reached 87% and 95% on the UCF-101 data set, respectively. Experiments show that the fusion network designed in this paper is effective.

The accuracy rates of SE-R3D-LSTM network and other networks on UCF-101 data set are compared, as shown in Table 3.



(a) Accuracy curve



(b) Loss function curve

FIGURE 12. SE-R3D-LSTM training process.

TABLE 3. Accuracy comparison.

Network	UCF-101 data set
CNN+LSTM	83.2%
CNN+ResNet+LSTM	84%
CNN+ResNet+SE+LSTM	90%
C3D[37]	82.3%
R3D	85%
SE-R3D	87%
SE-R3D-LSTM	95%

Since the CNN network can only extract spatial features, while LSTM can extract feature information in time dimension, spatial features and temporal features are used as classification basis to identify, which improves the effect of behavior recognition. The Resnet network can not only extract more abstract features, but also overcome the network degradation caused by the increasing number of network layers. Therefore, CNN- ResNet-LSTM network is 0.8% higher than CNN-LSTM in UCF-101 data set. Then, in order to improve the network performance, the attention mechanism SENet is introduced into the CNN-ResNet-LSTM network, making the recognition rate of CNN-ResNet-SE-LSTM on UCF-101 data set is 90%. This article uses the 3DCNN-LSTM fusion network. Firstly, the time domain

TABLE 4. SE-R3D-LSTM network parameters.

Network layer	Parameter quantity	Output size
SE-R3D	22502851	
Flatten		None×65536
Rshape		None×16×4096
LSTM	20975616	None×1024
Dense	103525	None×101

features can be extracted through 3DCNN to reduce the redundant information of feature vectors when sent to LSTM network. Secondly, due to the 4-dimensional input data of CNN, including the length, width, channel number and batch size, Batch size is the time dimension of the image sequence, so only one video sequence can be input at a time for training and testing. The input data of 3D-CNN is 5-Dimensional, and it can process multiple image frame sequences at the same time during training and testing, so the execution efficiency will be better. In this paper, an improved R3D network is proposed, which can reduce the dimension of feature map by changing the stride, improve the network efficiency, add batch normalization layer to improve the convergence speed of the network, and then add dropout layer to reduce the risk of over fitting. Therefore, the recognition rate of R3D network on UCF-101 data set is 2.7% higher than that of C3D network. In order to further enhance the effective features between layers, 3D-SENet network is introduced into R3D network. The recognition rate of SE-R3D network on UCF-101 data set is 86% higher than that of R3D network, which indicates that feature calibration is realized through 3D-Se module, and important features are given larger weight values, while useless eigenvalues are reduced. In order to fully extract the high-level temporal features, the LSTM is introduced into SE-R3D network, which makes the SE-R3D-LSTM 8% higher than the SE-R3D network in UCF-101 data set.

The description of SE-R3D-LSTM network framework designed in this paper is shown in Table 4.

It can be seen from Table 5 that the network parameters of SE-R3D-LSTM are 43581992, the network parameters of the convolutional layer are 22502851, accounting for 51.6% of the total parameters, the parameters of the LSTM layer account for 48.2% of the total parameters, and the final Dense layer only accounts for 0.2%.Through flatten layer, the output features of SE-R3D network are flattened. The content is to transform the multi-dimensional tensor of 3D convolution output into one-dimensional to 65536. The length of the video sequence becomes 16. After that, the ‘‘reshape’’ layer is used to increase the dimension of the input data vector, and the feature vector becomes 16 × 4096. Then, after passing through the GRU network, the output feature length is 1024. Finally, through the dense layer, softmax is used for classification.

Table 5 shows the comparison between the computational performance of the SE-R3D-LSTM network implemented in this article and other algorithms.

TABLE 5. Comparison of calculation speed between this algorithm and other algoritjms.

Algorithm	C3D	Two-Stream	3D-CNN	This article
Speed(fps)	313.9	12.5	603	900

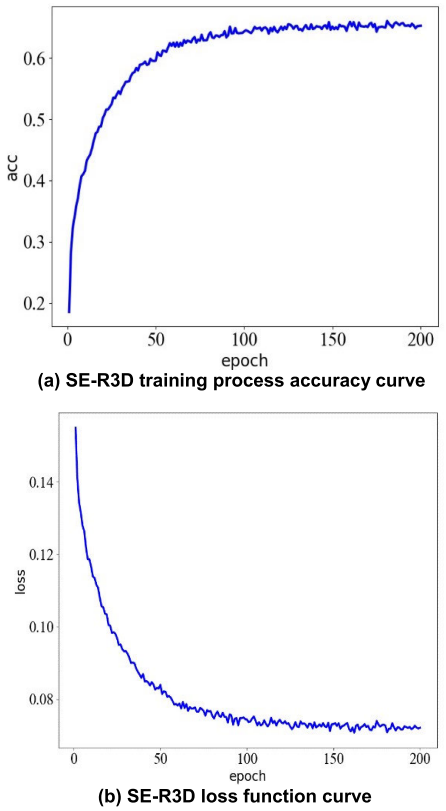
It can be seen from Table 6 that the calculation speed of the. C3D network is directly given in the literature, and other algorithms are obtained based on the algorithms given in the literature. For example, the Two-Stream algorithm uses two AlexNet networks, one AlexNet network uses optical flow for feature extraction, and the other AlexNet network uses RGB stream for feature extraction. The calculation speed of a single RGB stream is 25fps, and it takes 0.04 seconds to propagate once. The final calculation speed is 12.5 frames per second, which is much slower than that of the C3D network. The SE-R3D-LSTM network designed in this paper requires only 25ms to achieve a forward propagation, 16 frames of input images at a time, and 900 frames/s for calculation speed. Through continuous optimization of the network, although the recognition rate has not improved much, it has a greater advantage in computing speed.

In order to verify the effectiveness of the algorithm proposed in this paper, SE-R3D network and SE-R3D-LSTM network are tested on the HMDB-51 data set. The hardware platform used is the same as the one used above. The main hyperparameters required for the experiment are set as follows: the batch size is 16, the input video frame pixel is 32×32, the dropout is 0.2, the optimizer is SGD, the initial learning rate is 0.1, and the initial momentum coefficient is 0.9. The loss function of the network model is Softmax, which is the classification function. The training and loss curves of the SE-R3D network and SE-R3D-LSTM network on the HMDB-51 data set are shown in Fig.13 and Fig.14, respectively.

Fig.13 (a) is the SE-R3D training process accuracy curve, and Fig(b) is the SE-R3D loss function curve. As can be seen from Fig.13 (a), the network has a total of 200 epochs, and each epoch iterates 600 times. When the network was in the 100th epoch, the network convergence began to slow down, and in the 150th epoch, the network had basically converged, and the final accuracy rate of the SE-R3D network was 65.2%.

Fig.14(a) shows the accuracy curve of SE-R3D-LSTM training process, and Fig(b) shows the loss function curve of SE-R3D-LSTM. As can be seen from Fig.14(a), when epoch is about 200, the network begins to converge slowly. Until 250epoch, the recognition rate does not increase, and the network has tended to converge. Finally, the accuracy rate of SE-R3D-LSTM network is 69%.

Since the SE-R3D-LSTM network model is more complex and prone to over-fitting, this article uses the HMDB-51 and UCF-101 data sets. Although UCF-101 has more numbers than HMDB-51, the diversity of data is relatively poor, and the backgrounds of many actions are similar. In order to further improve the accuracy of SE-R3D-LSTM algorithm,



(a) SE-R3D training process accuracy curve

(b) SE-R3D loss function curve

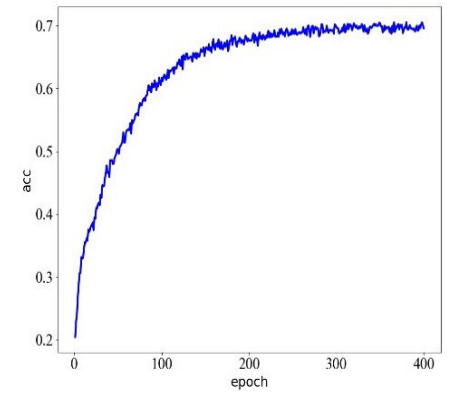
FIGURE 13. SE-R3D training process on HMDB51 data set.

TABLE 6. Accuracy of different algorithms on UCF-101 data set.

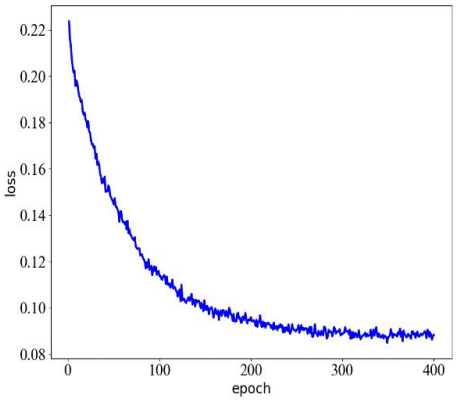
Network	UCF-101 data set
IDT+VideoLSTM	91.5%
P3D	88.6%
C3D	82.3%
Two-Stream-I3D	98%
LSTM+TWO-Stream[34]	90.1%
Se-R3D-LSTM	96.5%

the algorithm is pre trained with Kinetics data set, and then tested with ucf-101 data set. The specific results are shown in table 6.

Two-Stream-I3D achieves a 98% recognition rate on the UCF-101 data set. The specific steps of the algorithm are to merge the Two-Stream and InceptionV1 networks and expand them to three dimensions. Since the dimensionality on the timing sequence cannot be reduced too quickly, the convolution kernel before the fusion of the two networks is $1 \times 2 \times 2$, and the size of the convolution kernel after the fusion is $2 \times 2 \times 7$. The experimental results show that there is still a certain gap between SE-R3D-LSTM and Two-Stream-I3D algorithm, but for the algorithms that have been popular in the past two years, the SE-R3D-LSTM network designed in this article has a relatively large recognition rate. Literature [34] proposed three methods of TWO-Stream and LSTM network fusion. Finally, it was found that the recognition rate of the



(a) SE-R3D-LSTM training process accuracy curve



(b) SE-R3D-LSTM loss function curve

FIGURE 14. SE-R3D-LSTM training process on HMDB51 data set.

two networks without passing through the full connection layer was the highest, but TWO-Stream was not suitable for learning the information in a long video frame sequence. At the same time, the recognition rate of IDT + LSTM and P3D is also lower than that of SE-R3D-LSTM network, which proves the effectiveness of the network in behavior recognition

IV. CONCLUSION

Aiming at the low recognition rate of traditional algorithms, this paper proposes a behavior recognition algorithm of SE-R3D-LSTM. When a deep enough network is used, the representation ability and recognition rate of the network can be increased, but if it is only a single increase in the number of network layers, there will be problems such as vanishing gradient, so this paper proposes a ResNet network as the basic framework. Firstly, the ResNet module is extended to design the R3D network to make up for the ResNet network's inability to effectively process the low-level spatiotemporal features in the video, and to reduce the dimensionality of the feature map by reducing the stride to avoid using pooling layer, which not only improves the efficiency of the network, and to prevent the loss of useful feature information. At the same time, BN layer and Dropout layer are added, In order to improve the convergence speed of the network and control overfitting effectively. After that, because the global maximum eigenvalue also contains important global information, the GMP layer is added to the SENet network, and the global information in the spatiotemporal domain can be fully extracted after GAP and GMP operations. Then, in order to use the independent performance between convolution feature channels to better model, enhance useful features and weaken useless features, therefore, the SENet attention mechanism is introduced into the R3D network. After joining the LSTM network, the feature information of the time dimension can be better extracted. Finally, the accuracy of the SE-R3D-LSTM network on the UCF-101 data set reached 96.5%, which improved the behavior recognition rate to a certain extent.

Although the SE-R3D-LSTM network can effectively solve the problems of gradient disappearance and insufficient memory ability, but the R3D network has a lot of parameters, which increases the difficulty of network training, and the increase in the amount of calculation leads to overfitting during the training process,and it is easy to achieve the global optimal solution, resulting in no increase in recognition rate. Moreover, LSTM has more non-linear transformation calculations, which reduces the training speed of the network. For SE-R3D-LSTM network will have a large number of parameters, this paper puts forward the idea of optimizing the network: R3D network for feature extraction and compression in the time domain. The advantage of GAP network layer is that it doesn't need to optimize the output parameters of the last convolution layer, at the same time, it can reduce the huge amount of parameters and prevent over fitting, but it doesn't play much role in the time domain features, and the parameters will be solidified in each convolution layer of the network. Therefore, if we want to increase the recognition rate, we must remove the GAP layer and deepen the number of network layers. In the residual structure of R3D network, there are two convolution layers, and the convolution kernel is $3 \times 3 \times 3$. In this paper, $1 \times 1 \times 1$, $3 \times 3 \times 3$ and $1 \times 1 \times 1$ convolution kernels are used to replace the two $3 \times 3 \times 3$ convolution kernels. Two $1 \times 1 \times 1$ convolution layers are introduced, which not only increases the nonlinear transformation of the network, but also reduces the amount of parameters. At the same time, a convolution layer is added to make the network have better generalization ability.

REFERENCES

[1] D. Albright and G. R. Stroner, "Visual motion perception," *Proc. Nat. Acad. Sci. USA*, vol. 92, no. 7, pp. 2433–2440, 1995.

[2] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[4] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[6] D. Li, L. Yu, J. He, B. Sun, and F. Ge, "Action recognition based on multiple key motion history images," in *Proc. IEEE 13th Int. Conf. Signal Process. (ICSP)*, Nov. 2016, pp. 993–996.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[8] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.
[9] O. B. Ahmed, "SIFT Accordion: A space-time descriptor applied to human action recognition," in *Proc. Int. Conf. Mach. Vis.*, Nov. 2011, pp. 1368–1374.
[10] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.
[11] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[12] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
[13] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, pp. 561–580, 2007.
[14] A. Rosani and N. Conci, "Human behavior recognition using a context-free grammar," *J. Electron. Imag.*, vol. 23, no. 3, 2014, Art. no. 33016.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Comput. Linguistics*, vol. 1, no. 4, pp. 568–576, 2014.
[17] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 677–691.
[18] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[19] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497.
[20] A. Gavrilov, A. Jordache, M. Vasdani, and J. Deng, "Convolutional neural networks: Estimating relations in the ising model on overfitting," in *Proc. IEEE 17th Int. Conf. Cognit. Informat. Cognit. Comput.*, Berkeley, CA, USA, Jul. 2018, pp. 154–158.
[21] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 13, 2021, doi: 10.1109/TGRS.2021.3093334.
[22] X. Zhao, R. Tao, W. Li, H.-C. Li, Q. Du, W. Liao, and W. Philips, "Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7355–7370, Oct. 2020.
[23] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100–111, Jan. 2020.
[24] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Washington, DC, USA, Jun. 2016, pp. 770–778.
[26] A. Mahajan and S. Chaudhary, "Categorical image classification based on representational deep network (RESNET)," in *Proc. 3rd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Coimbatore, India, Jun. 2019, pp. 327–330.
[27] V. Atliha and D. Sesok, "Comparison of VGG and ResNet used as encoders for image captioning," in *Proc. IEEE Open Conf. Electr., Electron. Inf. Sci. (eStream)*, Vilnius, Lithuania, Apr. 2020, pp. 1–4.
[28] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[29] Y. Zheng, R. Wang, J. Yang, L. Xue, and M. Hu, "Principal characteristic networks for few-shot learning," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 563–573, Feb. 2019.
[30] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5534–5542.
[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[32] S. Khurram, R. Z. Amir, and S. Mubarak, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: https://arxiv.org/abs/1212.0402
[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[34] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 461–470.

JIN WU received the B.S. degree in automation major and the M.S. degree in control science and engineering from Xi'an Jiaotong University, in 1998 and 2001, respectively. She has been a Professor with the School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi, China, since 2014. She has a total of 30 articles. Her research interests include signal and information processing.

YI YUAN AN received the B.S. degree, in 2018. She is currently pursuing the M.S. degree with the Xi'an University of Posts and Telecommunications. Her research interests include image processing and behavior recognition.

QIAN WEN SHI was born in Baoji, Shaanxi, in 1996. She received the bachelor's degree in engineering, in 2018. She is currently pursuing the master's degree. She mainly studies the deep learning algorithm of microexpression recognition.

WEI DAI was born in Xi'an, Shaanxi, in 1996. He received the bachelor's degree in engineering, in 2019. He is currently pursuing the master's degree. He mainly studies behavior deep learning algorithms.