Journal of Electronic Imaging

JElectronicImaging.org

Spatiotemporal information deep fusion network with frame attention mechanism for video action recognition

Hongshi Ou
Jifeng Sun

Hongshi Ou, Jifeng Sun, "Spatiotemporal information deep fusion network with frame attention mechanism for video action recognition," *J. Electron. Imaging* **28**(2), 023009 (2019), doi: 10.1117/1.JEI.28.2.023009.

Spatiotemporal information deep fusion network with frame attention mechanism for video action recognition

Hongshi Ou* **and Jifeng Sun**
South China University of Technology, School of Electronic and Information Engineering, Guangzhou, China

Abstract. In the deep learning-based video action recognition, the function of the neural network is to acquire spatial information, motion information, and the associated information of the above two kinds of information over an uneven time span. We propose a network for extracting semantic information of video sequences based on the deep fusion feature of local spatial–temporal information. Convolutional neural networks (CNNs) are used in the network to extract local spatial information and local motion information, respectively. The spatial information is in three-dimensional convolution with the motion information of the corresponding time to obtain local spatial–temporal information at a certain moment. The local spatial–temporal information is then input into the long- and short-time memory (LSTM) to obtain the context relationship of the local spatial–temporal information in the long-time dimension. We add the ability of the regional attention mechanism of video frames in the neural network mechanism for obtaining context. That is, the last layer of convolutional layer spatial information and the first layer of the fully connected layer are, respectively, input into different LSTM networks, and the outputs of the two LSTMs at each time are merged again. This enables a fully connected layer that is rich in categorical information to provide a frame attention mechanism for the spatial information layer. Through the experiments on the three action recognition common experimental datasets UCF101, UCF11, and UCFSports, the spatial–temporal information deep fusion network proposed has a high correct recognition rate in the task of action recognition. [DOI: 10.1117/1.JEI.28.2.023009]

Keywords: video action recognition; spatial–temporal information deep fusion; frame attention mechanism; deep learning.

Paper 180837 received Sep. 25, 2018; accepted for publication Feb. 21, 2019; published online Mar. 12, 2019.

1 Introduction

Action recognition technology is a technology that the computer understands and classifies human action from an image sequence containing people. This paper uses deep learning methods to perform action recognition. Deep learning is a very popular direction in the field of machine learning in recent years. Convolutional neural networks (CNNs) have certain translation invariance and scale invariance, and their calculation methods have great similarities with mammalian visual systems. The CNN network has a significant improvement over the traditional neural network recognition, and the method is an end-to-end identification method that does not require manual design features, which has attracted a large number of people to research and has achieved success in many areas of computer vision.

In the deep learning-based video action recognition, the design of the neural network structure mainly focuses on how to obtain spatial information, motion information, and the association of the above two kinds of information over an uneven time span. Only by fully obtaining enough spatial information and motion information in the time dimension can the neural network better recognize the action. For example, when performing a motion recognition such as a swing in golf and a bicycle, the difference in spatial background information will be good classification identification information because in a swing of golf, the golf club, and golf ball are significantly recognized and most of the green background can be regarded as information of actional recognition. However, relying solely on spatial information is difficult to recognize actions with similar backgrounds, such as the golf swing action shown in Fig. 1 and the hitting action of the croquet sport. Because the backgrounds of these two actions are very similar, in the case of such a situation, the motion information clearly better recognizes the above two actions.

Recently, many researchers have had many outstanding achievements in such research: let single or multiple video frames enter single-channel CNN to learn local spatial–temporal information.[1–3] The CNN-acquired spatial information is merged with the motion information extracted by the traditional optical flow method through the CNN.[4–7] The CNN combines with the long- and short-time memory (LSTM) network to obtain video stream context information.[8–10] However, compared with the excellent application performance of CNNs in face recognition,[11] image classification,[12] and human pose estimation,[13] there is no substantial progress in the application of CNNs to action recognition. In fact, looking at the test performance on the actional recognition standard datasets UCF101 and HMDB51, among the current method of high video action recognition rate, in addition to single-channel CNN acquisition of recognition features through supervised learning,[14] there are methods for merging single-channel CNN learning extraction features with traditional artificial definition features (such as HOF[15]),[16,17,18] and the coding–decoding network architecture composed of CNN and LSTM[19,20,21] to implement action recognition.

Among the above methods, there are still some defects, mainly for the following reasons: (1) the training dataset for action recognition is relatively lacking. Different from image classification, there are large datasets like ImageNet (1000 samples per category) that can be used for training,

*Address all correspondence to Hongshi Ou, E-mail: ouhongshi@163.com

Fig. 1 Golf swing and croquet batting.

but the standard dataset for action recognition UCF101 has only 100 samples per action category. (2) The method of the multivideo frame simultaneously entering CNN to extract spatial–temporal information solves the problem of local spatial–temporal information fusion. However, because CNN does not get good information about video streams over time, such methods cannot solve the problem of full acquisition of time-dimensional (3-D) information. In addition, the dual-stream architecture method[22,23] based on spatial CNN network space features and temporal CNN network optical flow characteristics in full-connection layer fusion cannot obtain two important clues in video action recognition: (i) spatial location where motion occurs (ii) spatial information and changes in motion information over time. (3) In the CNN and LSTM methods, after the first fully connected layer of CNN is connected to the LSTM, although LSTM is used as a tool for 3-D information extraction since the fully connected layer has only semantic information, the time clue of spatial information is not obtained, so the regional attention mechanism of the video frame is also lacking.

Based on the above deficiencies, this paper proposes a neural network architecture that not only can fully integrate local spatial and motion information but also can obtain long-term information dimension and has added a video frame attention mechanism.

2 Related Work

In the recent research on CNN for video action recognition, the main problem is how to obtain the spatial information of video frames and the information over time. In order to obtain the spatial–temporal information of the video stream, the following three CNN-based network structures are proposed in Ref. 1: (i) late fusion: two video frames with a certain interval enter two different CNN channels, and finally merge at the first fully connected layer so that the actional motion information of the two frames before and after the acquisition can be obtained. (ii) Early fusion: allows multiple consecutive video frames to enter the single-channel CNN simultaneously. This approach is primarily concerned with obtaining more detailed motion information. (iii) Slow fusion (SFCNN): consecutive multiple video frames have overlapping into the four-channel CNN, and the four-channel CNN merges into two different CNN channels after the last convolution layer is merged. The dual channel CNN enters the single-channel CNN after the last convolutional layer is fused. SFCNN is also for considering how to obtain spatial–temporal information of video streams more effectively, so this network architecture allows more video frames to enter the network at the same time, but this also causes an increase in network parameters. The above method uses two-dimensional (2-D) CNN, i.e., the 2-D CNN acquires spatial features and then fuses to obtain motion information. In the method proposed in Refs. 2 and 16 consecutive frames in a finite time were entered into a CNN network (3DCNN) having a 3-D convolution kernel with a parameter of $3 \times 3 \times 3$. This method works because the convolution kernel acts on both the spatial dimension and the time dimension. Therefore, compared with the above-mentioned SFCNN, which only has a 2-D convolution kernel acting on the spatial dimension, the network architecture achieves better performance, but the network depth and parameters will be deeper

and more. For 3DCNN, Sun et al.[24] decomposed the 3-D volume integration into 2-D spatial convolution and one-dimensional (1-D) time convolution. The 1-D time convolution was a characteristic channel, in which 2-D space was convolved in time, and it was embedded only at the upper layers of the network. Simonyan and Zisserman[4] proposed a dual-flow structure model based on CNN, which used space CNN stream to extract spatial information of single video frame and used time stream CNN to extract motion information of multioptical stream frame. Then the extracted spatial information and the motion information were fused on the full link layer in the dual stream structure.

In addition to the above-mentioned spatial–temporal information using CNN to obtain video action, the combination of CNN and LSTM has attracted more and more attention from researchers. In the method of combining CNN and LSTM, the image features were first extracted using CNN, and the features extracted by CNN were sent to LSTM. For example, Zhu et al.[25] proposed the use of hypercolumn features for facial analysis. This so-called hypercolumn feature not only extracted the feature map of the last layer of the CNN as a feature of entering the LSTM but also extracted the features of the previous CNN layer and the late CNN layer into the LSTM. Kar et al.[14] made CNN's fully connected layer features into LSTM for time-dependent semantic information extraction. Among such methods, the feature map extracted by the previous CNN is rich in spatial features but lacks semantic information. In video action recognition, the most important thing is to extract a feature that is independent of position and rotation, but the features extracted by the previous CNN do not support this feature. On the contrary, in the features extracted by the late CNN full connection layer, the semantic information is rich but lacks spatial information.

Sharma et al.[26] and Yao et al.[27] proposed an LSTM-based attention model that added a mechanism to where actional classification of video frames should be addressed. Sharma et al.[26] used a soft focus mechanism—using the back propagation training method to dynamically change the region of interest of each frame in the video. This method of training focus coefficient weights using a backpropagation algorithm is a very resource intensive task, and the classification fails if there is an error in the area of interest. In addition, Karpathy et al.[1] proposed a multiresolution method with a fixed focus on the central portion of the video frame.

In the field of video recognition, there is a lack of label data for training networks. In order to overcome the shortcomings of the training samples, Ding et al.[28] designed a semisupervised deep domain adaptation framework for effective knowledge transfer, this paper's core idea is to jointly construct two coupled neural networks and build a classifier to enhance the feature transferability of the deep structure. Li et al.[29] proposed a transfer independent together method, which designed a general framework to solve the shortcomings of other methods that were only applicable to a particular situation or required target samples for training. Li et al.[30] proposed a low-rank discriminant embedding (LRDE) method. This paper focuses on the specific problem of multiview learning where samples have the same feature set but different probability distributions. LRDE not only deploys low-rank constraints on both the sample level and feature level to dig out the shared factors across different views but also preserves geometric information in both the ambient sample space and the embedding feature space by designing a graph structure under the framework of graph embedding. Li et al.[31] proposed a heterogeneous domain adaptation (HDA) method that can optimize both feature discrepancy and distribution divergence in a unified objective function. Specifically, they present progressive alignment, which first learns a new transferable feature space by dictionary-sharing coding and then aligns the distribution gaps on the new space. Different from previous HDA methods that are limited to specific scenarios, this approach can handle diverse features with arbitrary dimensions. Li et al.[32] proposed a multimanifold sparse graph embedding (MSGE) algorithm, which can explicitly capture multimodal multimanifold structure while considering both intraclass compactness and interclass separability and then learn an integral subspace model. Furthermore, a sparse embedding is achieved using L2,1-norm, which makes the transformation matrix row sparse, so MSGE can select relevant features and learn subspace transformation simultaneously.

Compared with the existing methods, the spatial–temporal information fusion network with attention mechanism proposed in this paper not only has the fusion mechanism of spatial information and motion information but also has the attention mechanism effectiveness of video frames.

3 Approach

In the action recognition neural network architecture proposed in this paper, video frame space information and time information are extracted using a pretrained convolutional network (CNN). The features from the spatial and temporal stream CNN are fused at the pixel level and then enter the LSTM for feature information extraction of the video series. The selection of CNN will be discussed in Sec. 3.1 of this paper. Section 3.2 discusses the fusion method of dual-stream CNN feature pixel level. Section 3.3 discusses the LSTM framework used by the action recognition model. Section 3.4 gives the question of how the characteristics of the dual-stream CNN fusion enter the LSTM and generate a mechanism of interest. Section 3.5 gives the overall network architecture. Section 3.6 introduces the specific implementation details of the network–network hyperparameter settings.

3.1 *Convolutional Neural Network Transfer Learning Implementation*

The proposed CNN in neural network models is used for feature extraction tasks. The traditional neural network learning method cannot obtain good feature extraction results under the limited training samples. Therefore, in order to improve the classification accuracy of the model, a large number of training datasets must be used for model training. In the action recognition task, the existing data are not very abundant, so if the model is trained from the original state, a good training model cannot be obtained. For the training problem of limited datasets, this paper uses the method of migration learning. In our network structure, CNN is used to extract the features of video frames. Inspired by the literature,[28–32] the form of our migration learning implementation is: the initial state of CNN is the VGG-16 model that has been pretrained using ImageNet dataset. The structure of the model is shown in Fig. 2. The VGG-16 model contains 13 layers

Fig. 2 VGG-16 network.

of convolutional layers and 3 layers of fully connected layers. The representation of the convolutional layer in Fig. 2 is conv<receptive_field_size><number_of_channels>, and the fully connected layer representation is FC-< number_of_channels>.

3.2 *Architectures for Fusing the Two Stream Networks*

There are two main drawbacks in the dual-stream fusion structure proposed in Ref. 4: (i) because fusion is only performed at the categorizable layer (FC layer), it is impossible to obtain information about the spatial and temporal characteristics at the pixel level through training. (ii) In the spatial information and time information extraction operation, the spatial CNN stream only works with a limited number of video frames, and the time CNN stream acts on a limited and fixed length video optical stream frame, which makes it difficult for the model to distinguish the action over time. The network structure of this paper is an extended version of the dual-flow structure. In order to overcome the two drawbacks of the dual-stream fusion structure, our network not only fuses at the FC layer but also fuses at the convolutional layer, which ensures spatial information and motion information fusion at the pixel level. In addition, the network uses LSTM to implement more frames into the network for both video frames and optical frames. This allows the network to better obtain the semantic information of the spatial–temporal features of video behavior. The final experiment proves that the neural network we designed can overcome the two drawbacks of the above dual-flow fusion structure.

3.2.1 *Spatial feature fusion method*

When distinguishing actions such as brushing the teeth and combing hair, the time stream CNN can recognize such motion because the hand has a periodic motion back and forth in space. The spatial stream CNN can recognize the position information of the motion (located in the teeth or hair), so the fusion of the two streams can identify whether to brush or comb the hair. For this reason, when the dual-stream network is fused, the fusion of the feature maps on each channel should be the corresponding fusion between pixels in the same position.

If the spatial stream CNN and the time stream CNN have the same network structure in the dual stream structure, the fusion method of the same spatial position on the feature map is very easy to implement, for example, it can be simply covered or stacked in the same location. However, each network (spatial CNN and time CNN) has multiple channels in the fusion layer. How to determine the channel correspondence between different networks is a key issue. We assume that different channels on the spatial network are responsible for extracting features from different regions of the space, whereas a channel in the time network is responsible for extracting motion features from different regions. Therefore, after the channels of the dual stream network are stacked on the fusion layer, the method of fusion must enable subsequent subnetworks to learn the ability to obtain a correspondence between one channel of one network and a channel of another network so as to better distinguish between different categories.

Based on the above considerations, we have chosen a spatial information fusion method based on convolution operations.[22] The mapping relationship of feature fusion is $y_t = f(x_t^a, x_t^b)$, where f is a fusion function, the spatial network feature map is $x_t^a \in \mathbb{R}^{H\times W\times D}$, the time network feature map is $x_t^b \in \mathbb{R}^{H'\times W'\times D'}$, and the merged output feature map is $y^t \in \mathbb{R}^{H''\times W''\times D''}$. In the above equation, t represents time (In the following, the operation is the same at each moment, so t is removed from each equation), w represents the width of the feature map, H represents the height of the feature map, and D represents the number of channels on the fusion layer. There are $\mathrm{H} = H' = H''$, $\mathrm{W} = W' = W''$, and $\mathrm{D} = D'$.

First, we make the feature maps of each channel on the fusion layer of the two networks in series, that is, have the following feature mapping relationship: $y^{\mathrm{cat}} = f^{\mathrm{cat}}(x^a, x^b)$, the specific stacking method is as follows:

$$y_{i,j,2d}^{\mathrm{cat}} = x_{i,j,d}^a \; y_{i,j,2d-1}^{\mathrm{cat}} = x_{i,j,d}^b, \tag{1}$$

where i and j are the feature map spatial position, d is the channel label, and there is $y \in \mathbb{R}^{H\times W\times 2d}$. The serial connection between the spatial network and the spatial network feature map does not define the correspondence between the dual-stream network channels. Therefore, the subsequent subnetworks need to develop the ability to learn the channel correspondence. Hence, we carry out convolution fusion on this basis:

$$y^{\mathrm{conv}} = y^{\mathrm{cat}} * f + b. \tag{2}$$

In the above equation, f is a set of filters and $f \in \mathbb{R}^{n\times m\times 2D\times D}$, $b \in \mathbb{R}^D$. The dimension of the filter is $n \times m \times 2D$, and the number of channels output after convolution fusion is D. The filter bank described above is used to reduce the dimension of the channel number and at the same time to fuse the feature maps x^a and x^b at the same spatial position. We set the parameters of f to be trainable, so f can learn the correspondence between the two feature maps. For example, if f is learned as a series of two permutation unit matrices $1' \in \mathbb{R}^{1\times 1\times D\times D}$, then the i'th channel of the spatial network is only summed with the i'th channel of the time network.

In the dual-stream network structure, another key issue in feature fusion is the feature fusion at the given level. As shown in Fig. 3, the method of fusion has many forms. The multilayer fusion method used in the right diagram of

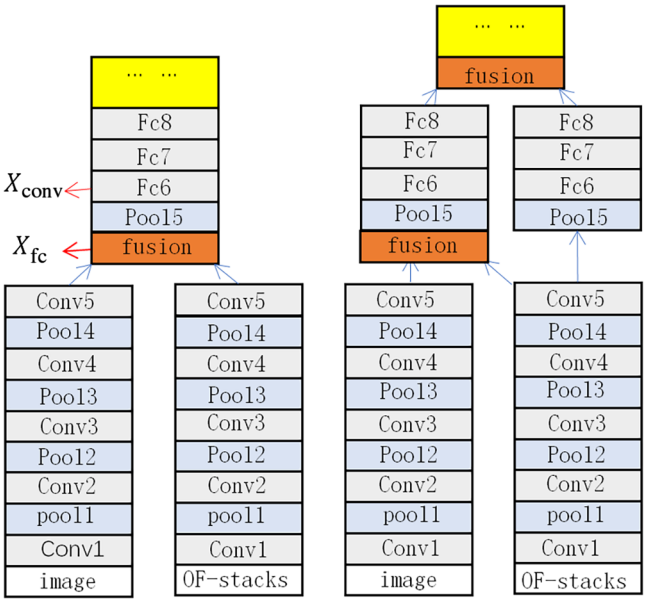

Fig. 3 Spatial information and motion information fusion method.

Fig. 3 is not only fusion in the spatially rich convolutional layer but also in the fully connected layer with rich semantic information. This fusion method has the pixel-level fusion of spatial information and motion information, and the complexity of semantic information is further improved, which enhances the robustness of network classification. However, the main parameters of the network come from the fully connected layer, so this fusion method greatly increases the number of network parameters. In this paper, the fusion method shown in the left side of Fig. 3 is used. This method only fuses in the convolutional layer with rich spatial information, so it does not increase the parameter amount of the network. Through postvalidation, this fusion method can achieve a good action recognition accuracy rate when combined with the network design method introduced in the following chapters.

3.3 *Layer Normalization LSTM Network*

In this paper, we need to obtain the correlation information of a sequence of frame images in the video for action category judgment. We use LSTM network to realize the feature extraction of video sequence signals. The "cell" structure of the traditional LSTM cycle network is shown in Fig. 4.

In order to obtain more abundant time information features, we make video frames continuously input into LSTM, and LSTM and CNN are combined to perform end-to-end training methods. Therefore, the depth of the network proposed in this paper is relatively deep, and the LSTM input gate i_t, the forgetting gate f_t, the activation function of the output gate o_t are all sigmoid functions. In order to ensure the convergence of the training, the LSTM cyclic network with layer normalization function is used in the network. That is, the parameters of each input sigmoid function are subjected to the following normalization process:

$$\mu_\beta \leftarrow \frac{1}{m}\sum x_i, \tag{3}$$

$$\sigma_\beta^2 \leftarrow \frac{1}{m}\sum (x_i - \mu_\beta)^2, \tag{4}$$

$$\widehat{x_i} \leftarrow \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}}, \tag{5}$$

$$y_i \leftarrow \gamma \widehat{x_i} + \beta \equiv BN_{\gamma,\beta(x_i)}. \tag{6}$$

In the above four equations, m is the number of samples used in single training, x_i is the i'th sample value in a single-training sample, and y_i is the sample value of the input sigmoid activation function after normalization.

3.4 *CNN and LSTM Combination Method for Action Recognition*

We use the combined method[33,34] shown in Fig. 5, where X_{conv} and X_{fc} specific network feature layers are shown in Fig. 3. X_{conv} is a feature map of the spatial and time networks after 3-D convolution fusion of the last convolutional layer, and X_{fc} is the first fully connected layer of the subnetwork behind the fusion layers.

In Fig. 5, the LN_LSTM* [layer normalization LSTM (LN_LSTM)] is a network as shown in Fig. 6(b), which is a cyclic network, in which the input is a time series signal and the output is also a time series. The LN_LSTM network is shown in Fig. 6(a) as a long-short-term memory network that is traditionally understood, i.e., sequence input and single-hidden layer unit output.

In the multilayer long-term memory network proposed in the paper, we first output the long- and short-term memory network (LN_LSTM*) through the sequence input sequence to realize the hidden sequence state expression of the video. Each hidden state in LN_LSTM* is input to the LN_LSTM

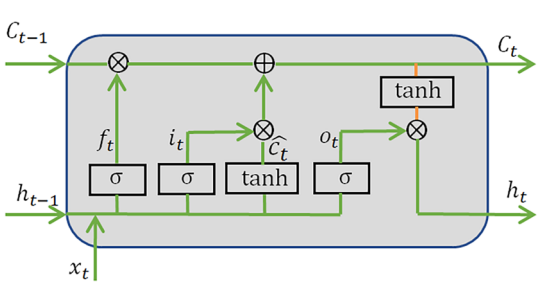

Fig. 4 LSTM cyclic neural network "cell" structure.

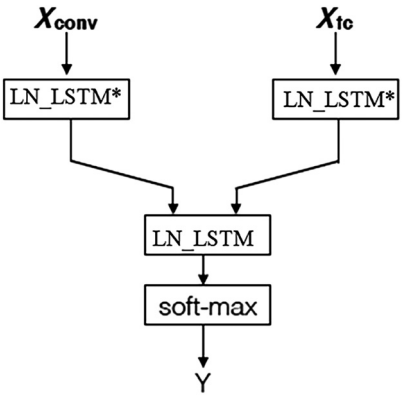

Fig. 5 CNN and LSTM combined network.

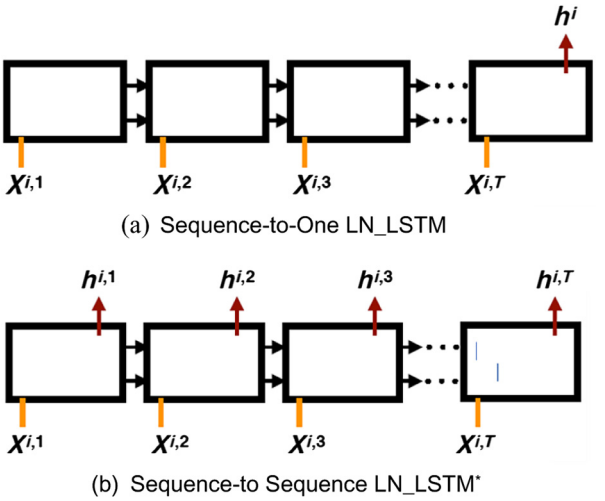

Fig. 6 Two forms of LSTM network used in the combined network: (a) sequence-to-one LN_LSTM and (b) sequence-to-sequence LN_LSTM*.

output by a single-hidden unit, and the final single-hidden unit expresses the video sequence. In this way, we achieve hierarchical acquisition of information, and through the end-to-end back propagation training method, the information exchange between the convolution spatial feature stream $X_{\text{conv}}-\text{LN\_LSTM}^*$ and the full-link semantic information stream $X_{\text{fc}}-\text{LN\_LSTM}^*$ is realized. After the semantic information and spatial information are exchanged, the video frame region attention mechanism of the network can be realized, and the back propagation training convergence process is also improved.

The network model equation is as follows. In the first layer, the LN_LSTM fused convolutional layer features and the fully connected layer features enter the long-term and short-term memory network of the "sequence input sequence output" mode of operation, realizing the sequence hidden state of the video:

$$h_{\text{conv}}^{i,t}=\text{LN\_LSTM}^*(x_{\text{conv}}^{i,t},h_{\text{conv}}^{i,t-1}), \tag{7}$$

$$h_{\text{conv}}^{i}=(h_{\text{conv}}^{i,1},h_{\text{conv}}^{i,2},\ldots,h_{\text{conv}}^{i,T}), \tag{8}$$

$$h_{\text{fc}}^{i,t}=\text{LN\_LSTM}^*(x_{\text{fc}}^{i,t},h_{\text{fc}}^{i,t-1}), \tag{9}$$

$$h_{\text{fc}}^{i}=(h_{\text{fc}}^{i,1},h_{\text{fc}}^{i,2},\ldots,h_{\text{fc}}^{i,T}). \tag{10}$$

In the second layer, the sequence hidden state of the LN_LSTM* output is re-entered into the LN_LSTM of single-sequence input hidden state output:

$$h^i=\text{LN\_LSTM}[W(h_{\text{conv}}^{i},h_{\text{fc}}^{i})], \tag{11}$$

$$y^i=\text{softmax}(h^i). \tag{12}$$

As shown in Fig. 7, the feature maps obtained when classifying the first, sixth, and ninth volumes of the space network (only the first six feature maps are shown in this figure): It can be known from the first-layer feature map that when the "shooting" is classified and learned, the "weight coefficient" of the spatial network is affected by the time network information in the process of learning acquisition, and finally the attention to the target is realized, that is, each edge of the target can be extracted correctly. In

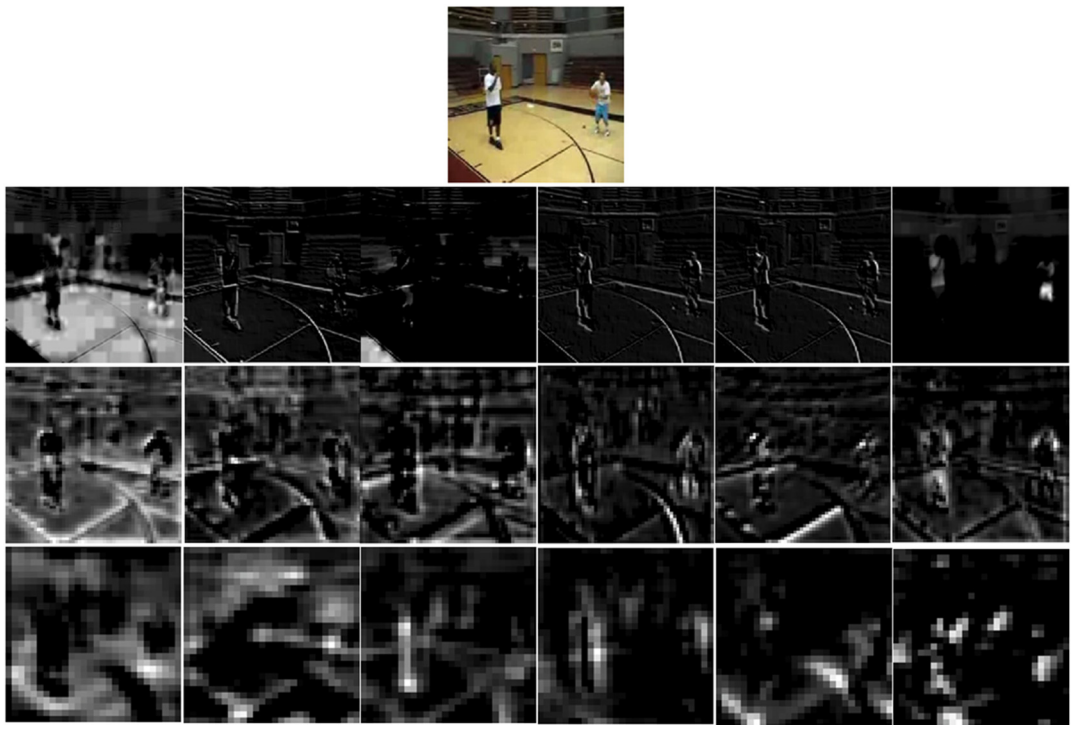

Fig. 7 Heat maps of Intermediate CNN layer outputs for the video frame obtained from UCFSports dataset: outputs are obtained for the first, sixth, and ninth convolutional layers and only the six channels for each layer is shown in this figure.

addition, the space network can also obtain various spatial information in the scene—the characteristics of the basketball court, that is, the network we designed can pay close attention to the target and obtain the scene information. The sixth and ninth layers realize the display. As the number of layers increases, the feature information of the network deep network is gradually simplified, and finally the output features are independent of the position and rotation. For this reason, we put the spatial information and time information into the final layer of convolution. The spatial–temporal information obtained after the fusion enters the following LSTM to obtain the semantic information of the video sequence.

3.5 *Proposed Architecture*

As described in Secs. 3.1–3.4, we will present our proposed network architecture as shown in Fig. 8. The spatial network and the time network perform 3-D convolution fusion on the last layer of convolutional layer (after ReLU output), and the fused feature output enters the sequence input sequence to output the LSTM network LN_LSTM*. At the same time, the feature given by the fusion is pooled and then enters the long-and short-term memory network LN_LSTM* of another sequence input sequence output. Finally, the sequence features of the two LN_LSTM* outputs are fused using the LN_LSTM output from the sequence input single-hidden layer unit, and the last layer of the network is the softmax classification layer.

In order to classify the video sequences, T frames are acquired in each video and enter the spatial network. The sampling time of the video frames shown in Fig. 8 is $t, t+\tau, \ldots, t+T\tau$, respectively. In addition, the input information of the network is centered on the information collection time of the space network, and the time characteristics are collected in the range of the center point time $L/2$. That is, the collected optical flow frame is collected centering on the captured video image frame. If $L = 4$, a total of the optical image of the $T \times 4(L)$ frame is collected into the time network. In addition, if $\tau \geq L$, the collection of optical flow frames will overlap.

3.6 *Implementation Details*

The spatial convolutional network and the time convolutional network we use are both VGG-16 network models. The VCC-16 model consists of 13 layers of convolutional layers. In the network structure of Fig. 7, the initial state of each convolutional layer of the spatial network is the pretrained VGG-16 model of the ImageNet dataset. The input of the spatial network is the video frame image. If we select the training dataset and the video with the smallest number of video frames with T frames, we set the number of sampling frames per video as T. The value of the number of frames T affects the validity of the semantic information acquired by the network. If the number of frames is too small, the correct rate given in Table 1 will not be achieved. The initial input of the time convolutional network is a multiframe stack of video optical flow images. The number of stacked frames we use is $L = 4$. The number L of stacks of optical flow frames also affects the correct rate. If L is <4, the change of frame motion before and after can not be obtained, which will result in the loss of motion information, which will reduce the correct recognition rate. Later experiments show that when $L = 4$, stepping can reduce the demand for GPU computing resources and at the same time can guarantee the high correct recognition rate as shown in Table 1. Before training, we will process the optical flow image of the video in advance, which can further shorten the training time. Since the number of channels of the first layer of convolutional layer of the input time convolutional network is related to the number of stacked frames of the optical stream image, the initialization of the first convolutional layer of the time network is random initialization, and the initial state of other convolutional layers is consistent with the initialization process of the spatial convolutional network, which is pretrained VGG-16 model for the ImageNet dataset. The size of the video frame image and the optical stream image of the input network are both 224×224.

As with the fusion structure described in Sec. 3.2 of this paper, the dimensionality of the 3-D convolution kernel f used in spatial–temporal information fusion is $3 \times 3 \times 3 \times 1024 \times 512$, where the dimension of the

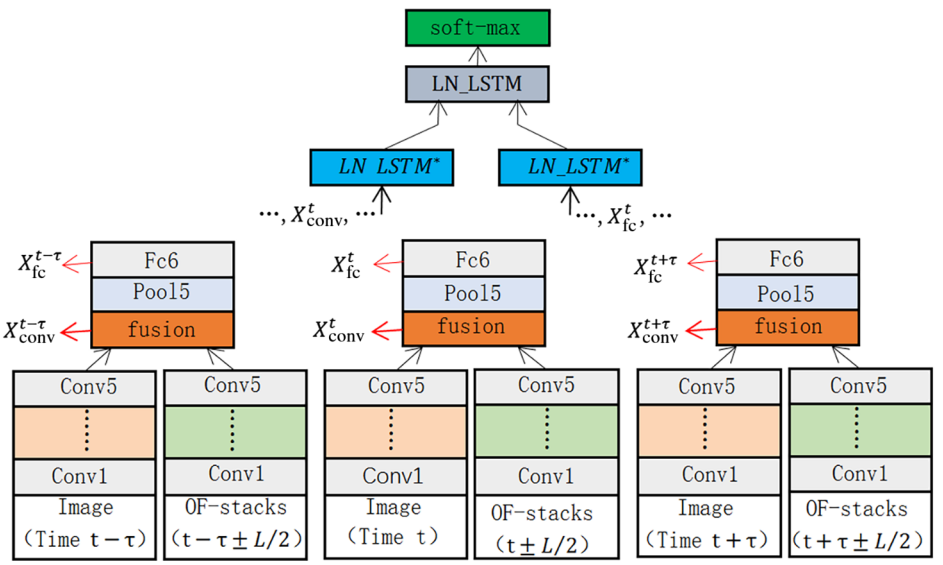

Fig. 8 Spatial–temporal information deep fusion network with video frame area attention mechanism.

Table 1 Comparison of our results to the state-of-the-arts on action recognition datasets UCF101.

Method	Accuracy (%)
LRCN[8]	82.9
Dense trajectories[35]	84.2
Composite LSTM model[20]	84.3
Soft attention[26]	84.9
C3D[2]	85.2
Two-stream ConvNet[4]	88.0
Beyond short snippets[21]	88.6
Snippets[36]	89.5
C3D + iDT + linear SVM[2]	90.4
Two-stream fusion[22]	92.5
Ours (VGG-16-3Dconv-2LST)	95.3

spatial–temporal filter is $H'' \times W'' \times T'' = 3 \times 3 \times 3$. $D = 1024$ is the number of channels after the feature map is stacked in the last convolutional layer (after the ReLU output) of the spatial convolutional network and the time convolutional network. $D' = 512$ is equal to the number of input channels of the next layer network (Fc6). In the LSTM network used in this paper, the hidden state dimension of all LSTM networks is 101. The tensor formats of the inputs X^t_{conv} and X^t_{fc} of LSTM* in Fig. 7 are $X^t_{\text{conv}} = (512 \times 14 \times 14)$, and $X^t_{\text{fc}} = (4096 \times 1)$, respectively.

At the same time, the spatial network, time network, and long-short-term memory are trained. The first fully connected layer (Fc6 in Fig. 8) in the network has a dropout ratio of 0.5 in training, and the output dropout ratio of all LSTMs is 0.25. Experiments show that if the dropout ratios of Fc6 are not equal to 0.5, or the dropout ratios of LSTM are >0.25, the correct rate of recognition will be reduced. The learning rate of the first training in network learning is 10^{-5}. When the network training converges and reaches a fixed classification accuracy rate, the training will stop. The second training uses the result of the first training as the initial state, which is the same as the subsequent training, and from the second training session, the initial learning rate is further reduced to 10^{-6}. The value of the learning rate will affect whether the training of the network can converge. The experiment proves that if the learning rate is $>10^{-5}$ in the first training, the network cannot converge.

Compared with the network with only dual-flow structure, the video frame region focusing mechanism of spatial–temporal information deep fusion network proposed in this paper will be overfitted. This is because the use of LSTM and the number of video frames sent to the network have an order of magnitude increase. To avoid overfitting and to enhance the generalization of the network, we use the method of increasing the training dataset to overcome the overfitting phenomenon. That is, the following random image frame acquisition process is added. In the selected T frame image, when the T frame is guaranteed to be acquired, the initial acquisition frame is randomly generated, and the acquisition interval τ is randomly sampled in the range.[2,10] In addition, the range of the sampling interval τ has an impact on the semantic information acquired by the network. If the interval is too large, the local semantic information is missing. If it is too small, the overall semantic information is missing. In addition, the value range of τ must contain the value of the number L of stacks of optical flow frames. If L is much larger than the maximum value of τ, local null-time information redundancy is caused. If L is less than the minimum value of τ, it causes the absence of local spatial–temporal information.

The computer used for network training is configured with two E5-2620 V4 processors and one NVIDIA GTX1080TI GPU. The computer system is a 64-bit UBUNTU system, and the software development environment is tensorflow. According to the above method, it takes 1 day (140 epoch) to train the network model for the first time to stabilize the network parameters. The second training takes only 8 h (46 epoch) to stabilize the network parameters.

4 Experiment

We use three public datasets to train and validate our network: UCF101, UCF11, and UCFSports. The above three datasets are very challenging for the recognition task of video action because the lighting conditions of each video are inconsistent. Moreover, in some video imaging processes, the camera does not move continuously, whereas in some video imaging, the camera does not move, and the background complexity of each video is different.

4.1 *Datasets*

1) UCF101:[37] UCF101 video dataset has 13,320 videos, including 101 categories.
2) UCF11: UCF11 video dataset has 1600 videos, including 11 action categories: ball-shooting, cycling, diving, golf swings, horseback riding, football kicking, swinging, tapping tennis, trampoline, volleyball smashing, and dog walks.
3) UCFSports: UCFSports video dataset with 150 videos, including 10 different action categories: diving, golf swing, kicking, weightlifting, horseback riding, running, skateboarding, swing bench, swing side, and walking. The video with the smallest number of frames in the dataset has 22 frames.

4.2 *Experimental Procedure*

Among the videos of each dataset, the video of the minimum number of frames has an inconsistent number of video frames. For each dataset, the minimum number of video frames T is selected as the number of frames of each video sample. Assuming that N is the total number of frames of a video sequence, the acquisition interval t is:

$$\mathrm{t} = \frac{N}{T}. \tag{13}$$

If the integer value of t is t', the sequence S of captured video frames is:

$$S = (1 \times t', 2 \times t', \ldots, T \times t'). \quad (14)$$

In this paper, the total video frames T sampled by the three datasets are all 17 frames. The experiment uses a cross-validation method, that is, each training dataset is randomly divided into a training set and a verification set. The training set is used to train the network model, and the verification set is used for the training correctness rate test verification after training. The ratio of the training set and the verification set is 7:3.

4.3 *Results and Comparative Analysis*

Figure. 9 shows a graph showing the relationship between the loss function and the number of echoes obtained by performing training verification on the UCF101 video dataset. It can be concluded from Fig. 9 that the action recognition network architecture training proposed in this paper can converge faster, and the correct recognition rate in the verification set tends to be stable with the increase of training echoes. From this, we can deduce that although our network depth is deeper than other methods, there is no case where the network gradient disappears or explodes.

In addition to the experimental verification of the video action recognition of UCF101 dataset using our proposed method, the UCF101 dataset is identified and classified by other methods. The experimental results of each method are shown in Table 1.

The LRCN[8] method caused the first fully connected layer and the second fully connected layer of the convolutional network to be sent to the LSTM for time information extraction, which lacked time change information extraction of spatial information. Dense trajectories[35] were a traditional dense trajectory method. The composite LSTM model[20] method used the video sequence frame as a sequence information with context semantics and used LSTM to implement sequence-to-sequence mapping for video characterization. This method does not use volume and neural network (CNN) to extract spatial information; C3D[2] and two-stream ConvNet[4] have the disadvantage of limited time dimension information. The beyond short snippets[21] method combines spatial and temporal information at the final predictive output layer, thus lacking pixel-level information fusion. The snippets[36] method implemented CNN to extract spatial features as a keyframe extraction tool and used the following SVM to classify keyframes. Such methods lack complete time dimension information. The two-stream fusion[22] method has the pixel level fusion of spatial information and mobile information but lacks the interaction process of spatial information and semantic information and lacks the ability to acquire complete semantic information. Combining the characteristics of each method and the analysis of the proposed network architecture, the proposed method has the ability to make up for various defects of other methods.

Figures 10 and 11 show graphs showing the relationship between the loss function and the number of training times for training verification of UCF11 and UCFSports video datasets. As illustrated in Figs. 10 and 11, the action recognition network architecture proposed in this paper is applied to the dataset with fewer samples, and the training convergence is faster. Moreover, in the case of limited training samples and verification sets, the correct rate is 100%, which makes it possible to judge that our network has a good generalization ability.

Similarly, in addition to the experimental verification of the video action recognition of UCF11 and UCFSports datasets using our proposed method, the UCF101 dataset is identified and classified by other methods. The experimental results of each method are shown in Table 2.

Cho et al.[38] used traditional multicore sparse representation to express video local motion features and global motion features. This method requires learning the dictionary to express local motion features and is limited by the dictionary's expressive ability; Weinzaepfel et al.[39] first used spatial CNN and time CNN to obtain multitarget candidate regions of frame images, racked regions with higher classification scores, and then evaluated the scores. Finally, the action was timed and classified using the form of a sliding window on the time dimension. Such methods lack the effective integration of spatial information and time information, so the classification results are improved with the relative

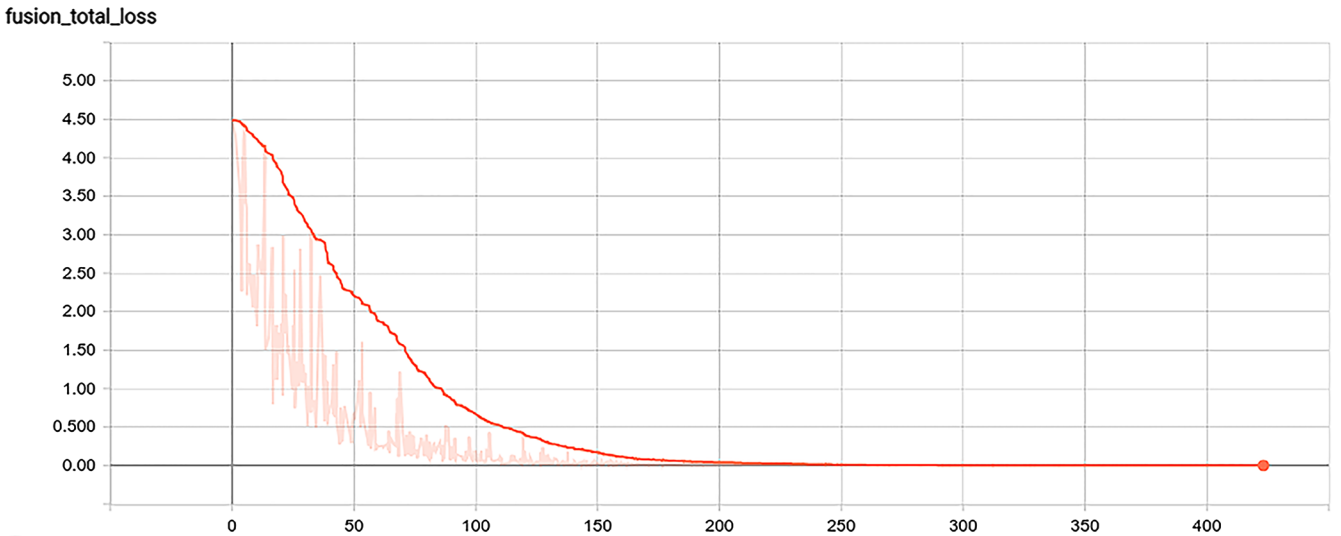

Fig. 9 The relationship between the loss function and the number of training obtained by performing training verification on the UCF101 video dataset.

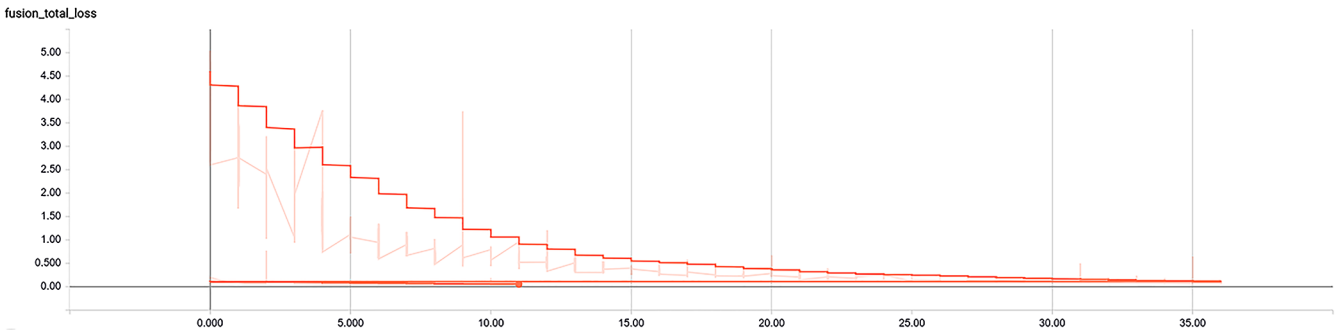

Fig. 10 The relationship between the loss function and the number of training obtained by performing training verification on the UCF11 video dataset.

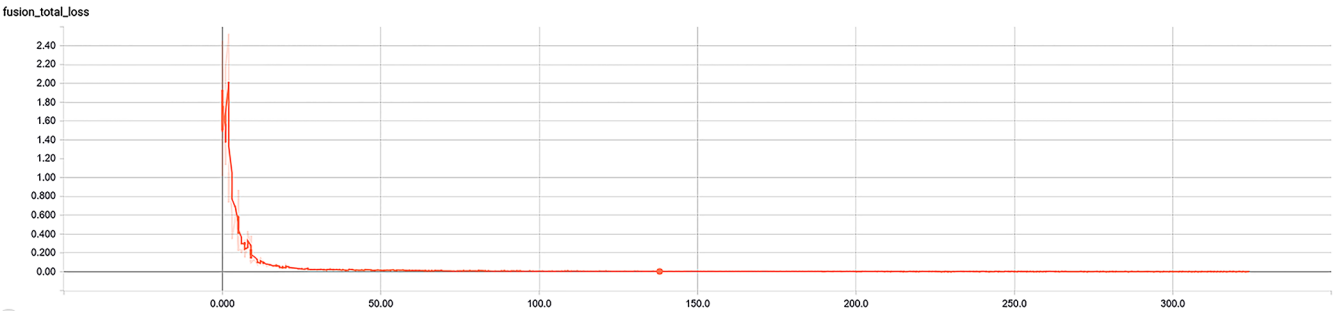

Fig. 11 The relationship between the loss function and the number of training obtained by performing training verification on the UCFSport video dataset.

Table 2 Comparison of our results to the state-of-the-arts on action recognition datasets UCF11 and UCFSport.

UCF11		UCFSport	
Method	Accuracy (%)	Method	Accuracy (%)
Dense trajectories[35]	84.2	Dense trajectories[35]	89.1
Soft attention[26]	84.9	Weinzaepfel et al.[39]	90.5
Cho et al.[38]	88.0	SGSH[40]	90.9
Snippets[36]	89.5	Snippets[36]	97.8
Two-stream LSTM[33]	94.6	Two-stream LSTM[33]	99.1
Ours (VGG-16-3Dconv-2LSTM)	100.0	Ours (VGG-16-3Dconv-2LSTM)	100.0

traditional methods, but there are still many areas for improvement; SGSH[40] first detected the video frame containing the target and the target in the corresponding frame image, extracted the frame containing the target, and described the global characteristics of the video sequence with histograms of oriented optical flow (HOOF) features. Meanwhile, the frame image target area was extracted, and the 3D-SIFT feature was used to describe the local target feature. Finally, the HOOF and 3D-SIFT features entered the multiclass SVM for class classification. This method is the best performing method in traditional nondeep learning methods because of the fusion of local features and global features. However, it is well-known that no matter what kind of artificially designed video global features and local features, it is impossible to contain more effective information than the neural network itself obtained through learning; Soft attention[18] had a frame attention mechanism, but it was not the full connection layer and the convolution layer working together to obtain spatial information. This method lacks the ability to express spatial information effectively, thus leading to errors in the classification of ball-shooting, throwing, and volleyball playing. Although two-stream LSTM[33] has a mechanism of attention, it lacks effective time information acquisition methods. The method proposed by our method is 100% accurate for UCF11 verification dataset. This proves that our network can effectively obtain motion information and spatial information, realize the effective fusion of spatial information and motion information, and finally obtain the semantic information of video.

Combining the characteristics of the above methods and comparing the action recognition network architecture proposed by our analysis, we can conclude that the network architecture designed in this paper can achieve the fusion of spatial–temporal information and the ability to express global features and local features. The experimental results also demonstrate the effectiveness of our method.

Disadvantages of our method: when the video classification is performed on the computer used in the network training, the specific prediction speed is slower—one video takes 770 ms. Among the various methods in Tables 1 and 2, their basic network structure is different. The experiment shows that the proposed method has better performance on the three datasets UCF101, UCFSport, and UCF11. However,

our experiments do not show that our method has the best recognition accuracy rate regardless of which dataset is used.

5 Conclusion

We propose a network for video action recognition. We use the migration learning method to fully acquire the video frame space information and the local time information of the video frame sequence using the pretrained CNN network with limited training samples. The local spatial information and the local time information are deeply fused, and finally the fused global time and space information is obtained using LSTM. In the design of LSTM, the fully connected layer semantic information is used to realize the effective extraction of the spatial information entering the LSTM (the spatial information attention mechanism is implemented). Through the experimental comparison on the standard dataset, the proposed method has a higher improvement in the correct recognition rate, which proves that we propose the network architecture with the following three important mechanisms: (1) the ability to fuse local spatial information with local time information, (2) effective acquisition of global spatial–temporal information, and (3) effective combination of spatial information and semantic information, and the effectiveness of spatial information attention mechanism.

Acknowledgments

Thanks to the support of the Doctoral Education Fund of School of Electronics and Information Engineering, South China University of Technology.

References

1. A. Karpathy et al., "Large-scale video classification with convolutional neural networks," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1725–1732 (2014).
2. D. Tran et al., "Learning spatiotemporal features with 3D convolutional networks," in *IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 4489–4497 (2015).
3. Z. Liu, C. Zhangb, and Y. Tianb, "3D-based deep convolutional neural network for action recognition with depth sequences," *Image Vision Comput.* **55**(2), 93–100 (2016).
4. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.* (2014).
5. L. Wang et al., "Towards good practices for very deep two-stream ConvNets," arXiv:1507.02159 (2015).
6. A. Diba, A. M. Pazandeh, and L. V. Gool, "Efficient two-stream motion and appearance 3D CNNs for video classification," in *European Conf. Comput. Vision* (2016).
7. Y. Zhu et al., "Hidden two-stream convolutional networks for action recognition," *CoRR*, arXiv:1704.00389 (2017).
8. J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2625–2634 (2015).
9. M. Xin et al., "ARCH: a daptive recurrent-convolutional hybrid networks for long-term action recognition," *Neurocomputing* **178**, 87–102 (2016).
10. C.-Y. Ma et al., "TS-LSTM and temporal-inception: exploiting spatio-temporal dynamics for activity recognition," *Sig. Proc. Image Comm.* **71**, 76–87 (2019).
11. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: a unified embedding for face recognition and clustering," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)* (2015).
12. C. Szegedy et al., "Going deeper with convolutions," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)* (2015).
13. J. Tompson et al., "Efficient object localization using convolutional networks," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)* (2015).
14. A. Kar et al., "AdaScan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *Conf. Comput. Vision and Pattern Recognit.*, (2017).
15. I. Laptev et al., "Learning realistic human actions from movies," in *IEEE Conf. Comput. Vision and Pattern Recognit.* (2008).
16. F. Zhu et al., "From handcrafted to learned representations for human action recognition: a survey," *Image Vision Comput.* **55**(2), 42–52 (2016).
17. L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deepconvolutional descriptors," in *IEEE Conf. Comput. Vision and Pattern Recognit.* (2015).
18. G. Chron, I. Laptev, and C. Schmid, "P-CNN: pose-based CNN features for action recognition," in *IEEE Int. Conf. Comput. Vision (ICCV)* (2015).
19. H. Chen et al., "Action recognition with temporal scale-invariant deep learning framework," *China Commun.* **14**(2), 163–172 (2017).
20. N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. 32nd Int. Conf. Mach. Learn.*, pp. 843–852 (2015).
21. J. Y. H. Ng et al., "Beyond short snippets: deep networks for video classification," in *2015 IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 4694–4702 (2015).
22. C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, arXiv:1604.06573 (2016).
23. Y. Wang et al., "Two-stream SR-CNNs for action recognition in videos," in *Br. Mach. Vision Conf.* (2016).
24. L. Sun et al., "Human action recognition using factorized spatio-temporal convolutional networks," in *IEEE Int. Conf. Comput. Vision (ICCV)* (2015).
25. C. Zhu et al., "Weakly supervised facial analysis with dense hypercolumn features," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR) Workshops* (2016).
26. S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," arXiv:1511.04119 (2015).
27. L. Yao et al., "Describing videos by exploiting temporal structure," in *IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 4507–4515 (2015).
28. Z. Ding, N. M. Nasrabadi, and Y. Fu, "Semi-supervised deep domain adaptation via coupled neural networks," *IEEE Trans. Image Process.* **27**(11), 5214–5224 (2018).
29. J. Li et al., "Transfer independently together: a generalized framework for domain adaptation," *IEEE Trans. Cybern.* 1–12 (2018).
30. J. Li et al., "Low-rank discriminant embedding for multiview learning," *IEEE Trans. Cybern.* **47**(11), 3516–3529 (2017).
31. J. Li et al., "Heterogeneous domain adaptation through progressive alignment," *IEEE Trans. Neural Networks Learn. Syst.*, 1–11 (2018).
32. J. Li et al., "Multi-manifold sparse graph embedding for multi-modal image classification," *Neurocomputing* **173**(3), 501–510 (2016).
33. H. Gammulle et al., "Two Stream LSTM: a deep fusion framework for human action recognition," in *IEEE Winter Conf. Appl. Comput. Vision* (2017).
34. Y. Xu et al., "DTA:Double LSTM with temporal-wise attention network for action recognition," in *3rd IEEE Int. Conf. Comput. and Commun.*, pp. 1676–1680 (2017).
35. H. Wang et al., "Action recognition by dense trajectories," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3169–3176 (2011).
36. M. Ravanbakhsh et al., "Action recognition with image based CNN features," *CoRR*, arXiv:1512.03980 (2015).
37. K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: a dataset of 101 human actions classes from videos in the wild," *CoRR*, arXiv:1212.0402 (2012).
38. J. Cho et al., "Robust action recognition using local motion and group sparsity," *Pattern Recognit.* **47**(5), 1813–1825 (2014).
39. P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *IEEE Int. Conf. Comput. Vision*, Santiago, Chile (2015).
40. A. Abdulmunem, Y.-K. Lai, and X. Sun, "Saliency guided local and global descriptors for effective action recognition," *Comput. Visual Media* **2**(1), 97–106 (2016).

Hongshi Ou obtained a master's degree in engineering from Chongqing University in July 2012 and is currently pursuing PhD in engineering at South China University of Technology. The research direction is action recognition and location.

Jifeng Sun is a senior member of the Chinese Institute of Electronics, a member of the IEEE Institute, and a member of the China Graphic Graphics Society. His research topic is machine learning.