

Action recognition based on two-stream spatio-temporal residual attention network

Jingjing Han¹, Hua Huo²

¹ Henan University of Science and Technology, School of Information Engineering, Luoyang 471003, China

² Henan University of Science and Technology, Engineering Technology Research Center of Big Data and Computational Intelligence, Luoyang 471003, China

Abstract: Aiming at the problems that there are few video action samples in the human action recognition process, overfitting in the training process, poor extraction ability of network models for key features, and low recognition accuracy. In this paper, we propose a method of action recognition based on a two-stream spatio-temporal residual attention network. First of all, by adding the Convolutional Block Attention Module (CBAM), we strengthen the model's channel and spatial feature extraction of video and suppress interfering information. Based on the output of each stream, train two convolutional neural networks to model static spatial information and motion information respectively, in order to combine the outputs of the two networks, we use weighted fusion to generate the final action recognition results. A large number of experiments were conducted in UCF101 and HMDB51 datasets, and the recognition accuracy reached 92.9% and 73.66%, respectively. The experimental results showed that the model can make full use of the spatio-temporal information in the video and extract the key information of motion.

Keywords: Action recognition; Attention mechanism; Two-stream network; Feature fusion.

1. Introduction

With the rapid development of mobile Internet and big data and other related technologies, video has become one of the main carriers of information. Facing the increasing video data, how to analyze and understand the content of the video is crucial, human action recognition as an important part of video understanding, has attracted more and more people's attention. Human action recognition mainly refers to the use of corresponding computer technology to process the collected video data, so as to be able to distinguish the type of human action. As a basic and extremely challenging task of computer vision, human action recognition has broad application prospects in many fields such as smart home [1], intelligent security [2], human-computer interaction [3], video retrieval [4] and so on. With the gradual strengthening of computer performance, the research content of action recognition has also developed to a deeper level, from the previous single background to today's complex background, from simple actions to complex actions, from single actions to continuous actions, these changes put forward higher requirements for human action recognition in video.

With the successful application of deep learning in image classification and object detection, many researchers have also applied it to human action recognition, unlike images that only have static visual appearance information, videos also contain time motion, which requires network models to extract both static visual information and motion information. For example, 'diving' action video usually involves a series of actions, such as 'jumping off the platform', 'spinning in the air' and 'falling into the water'. Secondly, for the human action in the video, the action of some parts is more important than that of other parts. For example, in the process of playing the violin, the human body's torso is still in the state of motion during the whole action. Our attention should focus on the movement of the object's hand. Therefore, how to assign different weights to action features of different locations is a challenging task. Given the above limitations, we propose a

multi-stream method of deep neural network, namely spatio-temporal residual attention network (STRA-Net), which uses multimodal information for video action classification. Our contributions are summarized as follows :

- 1). The proposed two-stream network model of spatio-temporal residual attention can make full use of temporal and spatial features to improve the accuracy of action recognition.
- 2). All video data are preprocessed by the data augmentation method to reduce the over-fitting caused by fewer samples.
- 3). The convolution block attention module (CBAM) is added to the network model to focus on the key area information and important channel information of the video, which makes the extracted features more distinguishable and enhances the expression ability of the features.
- 4). Using a weighted fusion method to combine the outputs of each network, it has been experimentally proved that when the RGB:Flow is 0.35:0.65, the model has the strongest recognition ability.

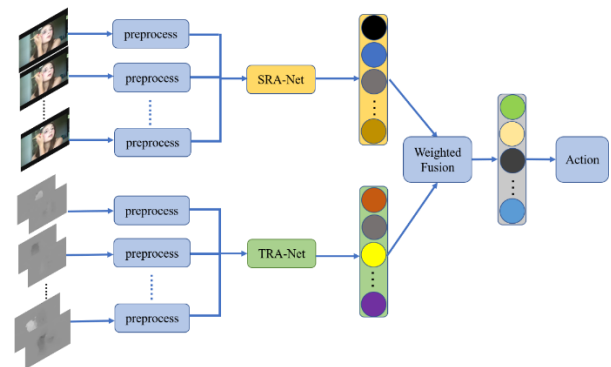


Fig. 1 The overall framework of spatiotemporal residual attention network

2. Related work

The traditional methods of action recognition generally use dense trajectories method[5], Space-Time Interest Points

method[6], Histogram of Oriented Optical Flow[7] and motion history image method[8] to extract features. Then the extracted features are encoded, and finally the traditional classifiers such as KNN and SVM are used for action classification. However, in the process of feature extraction and coding, due to a large amount of video data and the difficulty of feature extraction, it often leads to high spatiotemporal complexity, so it is generally only suitable for specific and small-scale research tasks. With the rapid development of deep learning methods, major breakthroughs have been made in tasks such as target detection, image recognition and image segmentation. Researchers have begun to apply deep learning methods to action recognition problems. Firstly, the original video is input into the deep network to automatically learn features, and then the classification results are output. In order to realize the effective modeling of video spatiotemporal information, many different types of methods are proposed, which can be divided into three categories: methods based on two-stream networks, methods based on recurrent neural networks and methods based on 3D convolutional networks.

2.1. Methods based on two-stream network

In order to model the spatiotemporal information in the video, the two-stream network uses two parallel networks to extract temporal features and spatial features respectively. By fusing the softmax scores of the two streams, the final prediction results of the video are obtained. The representative two-stream methods are Two-stream [9], TSN [10] and FlowNet [11]. The limitations of the two-stream method : (1) The two-stream network requires optical flow features, and the optical flow only represents the motion information of two adjacent frames, and the access to the time context is very limited, resulting in a small time span of the extracted action features. (2) Optical flow can only be extracted manually, and the scale of datasets in the field of action recognition is expanding, which makes the computational cost and storage space of optical flow larger and larger.

2.2. Methods based on recurrent neural network

For the weak modeling ability of two-stream networks for long time, recurrent neural networks (RNN) can be used to obtain temporal correlation features, and the commonly used LSTM [12] can effectively solve the problem of gradient dispersion as well as gradient explosion in RNN, so that the length of the input video is no longer limited and can better capture the interdependent features among long-time video data. Ng et al [13] combined convolutional neural networks with recurrent neural networks (CNN-LSTM) together to characterize video data in both spatial and temporal dimensions simultaneously. However, most LSTM networks are directly fed with high-dimensional features obtained from CNN fully connected layers, missing features in terms of action details. On this basis Ballas et al [14] proposed multilayer Gated-Recurrent-Unit Recurrent Networks (GRUs) to elucidate the complementarity between using low-level perceptual features and high-level perceptual features. To demonstrate that hierarchical feature fusion methods can improve recognition accuracy, Gammulle et al [15] used multiple feature fusion strategies to input features from different levels of the CNN into the RNN network. Since it is easy to lose the correlation of spatial locations by vectorizing

the corresponding spatial features directly at the corresponding spatial locations and using them as input to the LSTM, Li et al [16] used a convolution operation instead of matrix multiplication in the LSTM unit, which can maintain the spatial structure of the input data and achieve further encoding of spatial information. donahue et al [17] proposed the LRCN (Long-term Recurrent Convolutional Network) model to achieve effective extraction of video time series by extracting features from individual video images and inputting them into the corresponding LSTM to achieve the classification of different videos.

2.3. Methods based on 3D convolutional network.

Video is a time series composed of a large number of different and complex frames, which leads to the difficulty of spatio-temporal representation. Different from the two-stream network, the 3D convolutional network does not need to extract flow manually. It can directly input the video and use 3DConv to extract the spatiotemporal information of the video. Such as C3D[18], Res3D[19] and I3D[20]. These methods have complex network structures and need to be trained using very large datasets to get good results. To reduce the complexity of training 3D networks, many decomposition methods have been proposed, such as P3D[21], R(2+1)D[22], FstCN[23] and MiCTNet[24]. Among them, P3D and R(2+1)D split the 3D convolution kernel into a 2D spatial convolution and a 1D temporal convolution, while FstCN and MiCTNet use both 2D convolution and 3D convolution in a single network. Limitations of 3D network: a large number of parameters and complex network; only relatively short time intervals can be considered, and long-term time information cannot be obtained.

2.4. Attention mechanism

Attention mechanism is based on the use of the human visual mechanism. Human vision does not capture the information of the whole scene at once, but first selectively captures the important part, and then other parts. In the process of action recognition, we not only need to focus on the behavior of people in the video, but also need to capture and focus on the representative objects that can provide information. At present, many networks with attention mechanism are used for visual tasks. Focusing on the effective area can achieve better results. Hu [25] proposed the squeeze-and-excitation networks (SENet), which represents the correlation between the channel and key information by imposing a weight on the feature map of each channel. Zhang et al. [26] proposed the spatial self-attention mechanism, which makes the network pay more attention to the regions of interest. JADERBERG et al. [27] proposed an attention module to extract the key information in the image through spatial transformation of the image.

3. Proposed method

3.1. Spatio-temporal residual attention network (STRA-Net)

In this paper, we propose an action recognition network of spatio-temporal residual attention (STRA-Net), which is composed of a spatial residual attention network (SRA-Net) and a temporal residual attention network (TRA-Net). The composition structure is shown in Fig. 1. Both TRA-Net and SRA-Net use ResNet50 as the backbone network. The spatial

residual attention network learns the spatial characteristics and appearance characteristics of the video, and the temporal residual attention network learns the temporal information between the frames of the image. The two branches conduct independent predictions respectively. Finally, the prediction results of the two branches are weighted and fused to obtain the final action recognition results.

3.2. Convolution block attention module

The attention mechanism can be regarded as a resource allocation mechanism, which redistributes resources according to the importance of the attention object for the resources originally allocated equally. In the structural design of deep neural networks, the resource allocated by the attention mechanism is weight. CBAM consists of channel attention and spatial attention, as shown in Fig. 2. Spatial attention is to learn a weight matrix from the feature map of $H \times W$ size on the two-dimensional plane for all channels, and each pixel will learn a weight. These weights represent the importance of spatial location information. Different from spatial domain attention, channel domain attention imposes a weight on the feature map on the channel to represent the correlation between the channel and key information. The larger the weight is, the higher the correlation is. According to this correlation, the useful features are enhanced and the features that are not useful for current tasks are suppressed, so that the network can selectively enhance the useful feature channels and suppress the useless feature channels by using global information, so as to realize the adaptive calibration of feature channels. Given the initial feature map $F \in \mathbb{R}^{H \times W \times C}$ as input, the attention module calculates the channel attention map $M_c \in \mathbb{R}^{C \times 1 \times 1}$ and the spatial attention map $M_s \in \mathbb{R}^{1 \times H \times W}$ in turn. CBAM multiplies the input feature F and the channel attention weighting coefficient M_c to obtain a new feature F' . Then F' is multiplied by the spatial attention weighting coefficient M_s to obtain the final attention feature F'' . The calculation formula is shown in Equation (1):

$$F' = M_c(F) \otimes F, F'' = M_s(F') \otimes F' \quad (1)$$

Channel attention: The average pooling and the max pooling are used to compress the input feature map at the same time, and two different spatial context descriptors are generated: F_{avg}^c and F_{max}^c . Then, they are input into a multi-layer perceptron MLP for dimensionality reduction and dimensionality escalation operation. Finally, the channel attention map $M_c \in \mathbb{R}^{C \times 1 \times 1}$ is obtained through summation operation and Sigmoid function, as shown in Equation (2):

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ = \sigma(W_1(W_0(F_{avg}^c)) + W_1(W_0(F_{max}^c))) \quad (2)$$

Where $\sigma(\bullet)$ is the Sigmoid function, W_0 and W_1 are the weight parameters of the multi-layer perceptron?

Spatial attention : Two feature maps $F_{avg}^s \in \mathbb{R}^{1 \times H \times W}$ and $F_{max}^s \in \mathbb{R}^{1 \times H \times W}$ are obtained through average pooling and maximum pooling, and the two feature maps are cascaded and convolution operation is performed to obtain spatial attention map $M_s \in \mathbb{R}^{1 \times H \times W}$.

$$M_s(F) = \sigma(f([AvgPool(F'); MaxPool(F')])) \\ = \sigma(f([F_{avg}^s; F_{max}^s])) \quad (3)$$

where $\sigma(\bullet)$ is the Sigmoid function and $f(\bullet)$ is the convolution operation?

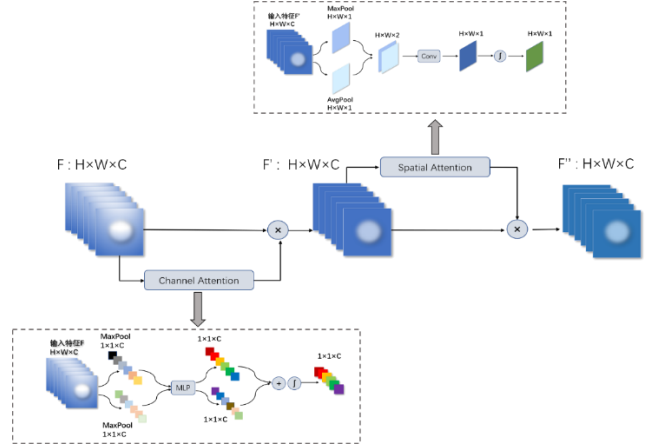


Fig. 2 The basic structure of CBAM

3.3. Construction of residual attention module

For neural networks, it is difficult to identify useful information. In order to extract more discriminative features in the image, CBAM is added to focus on the key area information and important channel information of the frame, as shown in Fig. 3. Firstly, the key information in the input feature is extracted by CBAM, and then the extracted key information is fused with the result of shortcut connections, and then added to the result of the residual part as the output feature of the whole module. Given that the input feature is x_i , the output result x_{i+1} can be expressed as:

$$H(x_i) = W'_i(x_i + A(x_i)) \quad (4)$$

$$x_{i+1} = H(x_i) + F(x_i, W_i) \quad (5)$$

where W'_i is a 1×1 convolution operation, W_i is a convolution operation, and $F(x_i, W_i)$ is an operation of the residual part.

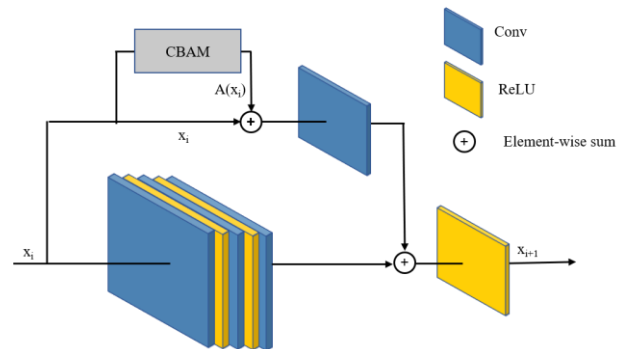


Fig. 3 Residual attention module structure.

3.4. Temporal residual attention network (TRA-NET)

The proposed method uses a 50-layer residual network (ResNet 50) as the backbone network after pre-training in the ImageNet dataset, and integrates the proposed residual attention module into ResNet-50. The structure diagram is shown in Fig. 4. RAM represents the residual attention module, and RM represents the residual module. TRA-NET

and SRA-NET networks have the same structure, except that the input data modalities are different during training.

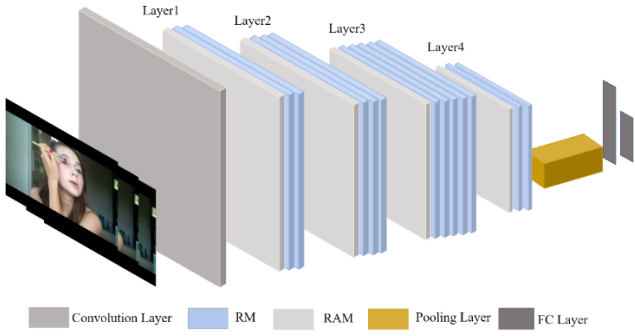


Fig. 4 Spatial residual attention network structure diagram

4. Experiment

4.1. Datasets

The method proposed in this paper is evaluated on two publicly available human action datasets: UCF101 and HMDB51. The video in the UCF101 dataset comes from YouTube, which contains 101 action categories. Each action category is recorded by 25 people and contains 13320 videos. The dataset contains rich types of actions, including body motion, human-human interactions, human-object interactions, playing musical instruments and sports. Its videos have great diversity, including camera motion, object appearance and pose difference, object scale and perspective change, cluttered background and uneven illumination. HMDB51 dataset video samples mainly come from movie shots and Google videos, including 51 action classes, a total of 6676 videos, each action class has more than 100 action clips. The whole action category is roughly divided into five categories: facial actions, body movements, interactive action with the object, facial actions with object manipulation and human action.

4.2. Implementation details

In the training stage, in order to make each branch network fully trained, the temporal stream and spatial stream network

are trained separately, and then the outputs of the two streams are weighted fusion to obtain the final action recognition results. In this paper, the accuracy of action recognition is used as the main evaluation criterion to measure the recognition ability of the model. Before training, in order to prevent overfitting, we use data augmentation to preprocess all videos. The size of each frame in the video is adjusted to 256×340 , then cut to 224×224 , and randomly cut and flip it. Finally, in order to further reduce the risk of network overfitting, dropout is used in all FC layers, that is, the nodes in the FC layer are randomly inactivated according to a certain probability. The experimental parameters are set in Table 1.

Table 1. Experimental parameters

parameter	value
learning_rate	1×10^{-3}
dropout	0.5
momentum	0.9
loss_function	Cross Entropy Los
optimizer	SGD

4.3. Experimental process

In order to obtain a better generalization ability of the model, the network is trained in three modalities: RGB, optical flow and RGB + optical flow. The three input modalities correspond to three networks: SRA-Net, TRA-Net and STRA-Net.

4.3.1. Training process

To better analyze their convergence process, the accuracy and loss curves of TRA-Net and SRA-Net on the UCF101 dataset are plotted respectively, as shown in Fig. 5. The blue line represents the accuracy, and the orange line represents the loss value. The initial learning rate is set to 0.001. When training SRA-Net, when the iteration is carried out to 6571 times, the learning rate is reduced to $1 / 10$ of the original. When the iteration is carried out to 13141 times, the learning rate is reduced by $1 / 10$ again. When training TRA-Net, every 8760 iterations, down to the original $1 / 10$, when the network iterations to 21900, the network stops training.

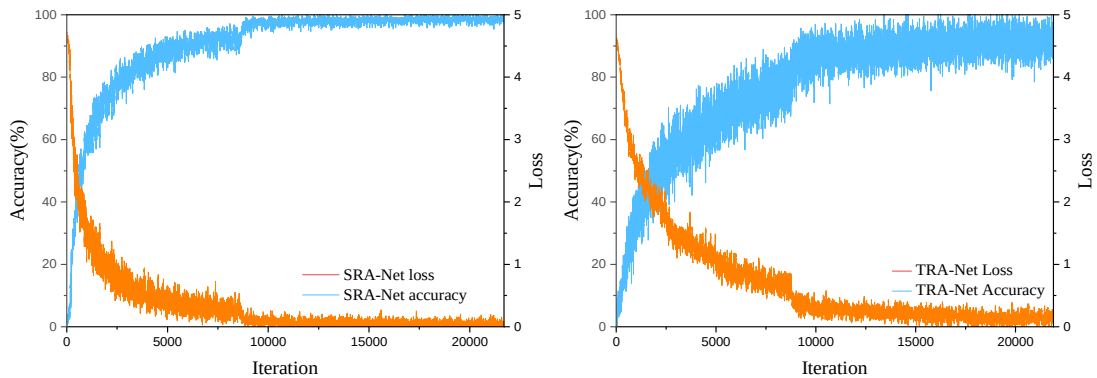


Fig. 5 Loss Value Curve and Accuracy Curve for SRA-Net and TRA-Net Training

It can be seen from Fig. 5, the accuracy and loss values of the above two networks fluctuate greatly in the initial iterative training, and after about 10,000 iterations, the accuracy and loss values tend to be stable.

4.3.2. Experimental results and analysis

This experiment analyzes and compares the accuracy changes of TRA-Net, SRA-Net and baseline networks on UCF101 and HMDB51 datasets. As shown in Fig. 6, the

above figure is the change process of accuracy on the UCF101 dataset, and the following figure is the change process of accuracy on the HMDB51 dataset. Among them, the black line represents TRA-Net, the red line represents SRA-Net, the blue line represents the baseline network with the input modality of optical flow, and the green line represents the baseline network with the input modality of RGB. It can be seen from Fig. 6 that the action recognition method of spatio-temporal residual attention is superior to the baseline network

in recognition accuracy. After obtaining the classification models TRA-Net and SRA-Net trained by optical flow and RGB frame data, the classification results of the two models are weighted and fused on the HMDB51 dataset in this experiment. As shown in Fig. 7, the horizontal ordinate represents the proportion of SRA-Net network, and the

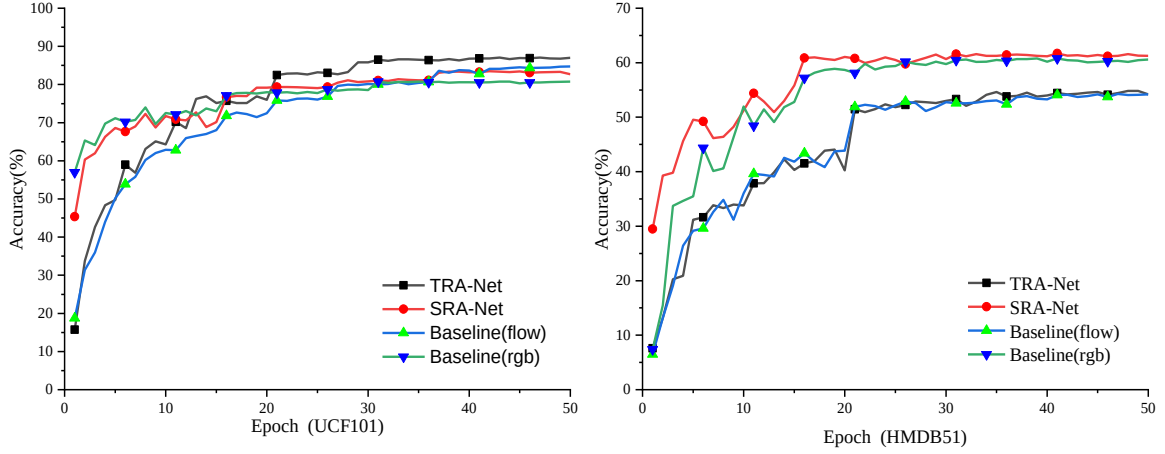


Fig. 6 Comparison with baseline networks on UCF101 and HMDB51 datasets

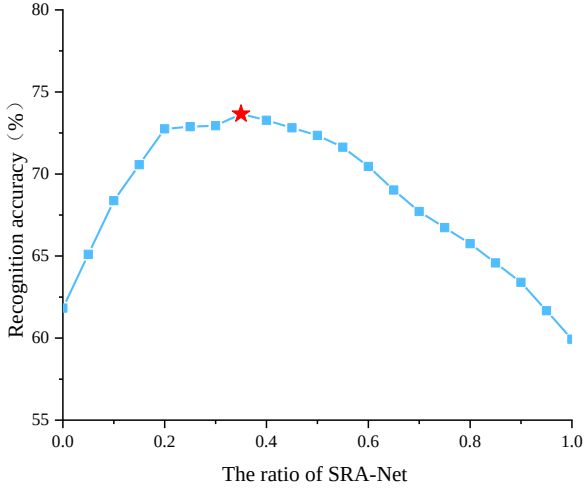


Fig. 7 Recognition accuracy under different weight ratios

4.3.3. Ablation experiment

In order to verify the effectiveness of the residual attention module, the proposed method is compared with the baseline network, and different input modalities are used to verify the proposed method. The comparison results are shown in Table 2. It can be seen that when the input modality is RGB, compared with the baseline network, the accuracy of the action recognition network of spatio-temporal residual attention proposed in this paper on the datasets UCF101 and HMDB51 is about 1.64 % and 1.79 % higher, respectively. When the input modality is Flow, compared with the baseline network, the accuracy of the action recognition network with spatio-temporal residual attention proposed in this paper on the datasets UCF101 and HMDB51 is about 2.58 % and 0.91 % higher, respectively. When the input modality is RGB + Flow, compared with the baseline network, the accuracy of the action recognition network with spatio-temporal residual attention proposed in this paper on the dataset UCF101 and HMDB51 is about 5.34 % and 2.26 % higher, respectively. It can also be found that the RGB + Flow input modality has achieved the best results on both datasets, which may be due to the complementarity between different forms of data, so that the model can learn better feature representation.

vertical ordinate represents the recognition accuracy of dataset video. When the weight ratio of SRA-Net and TRA-Net is 0.35: 0.65, the recognition ability of the whole model is the best. Therefore, this ratio is used for weighted fusion in subsequent experiments.

Table 2. Comparison of the recognition accuracy of the network under different input modalities

Network	Input Modalities	UCF101	HMDB51
Baseline	RGB	80.73%	58.73%
Baseline	Flow	84.65%	60.92%
Baseline	RGB+ Flow	87.56%	71.40%
TRA-Net	RGB	82.37%	60.52%
SRA-Net	Flow	87.23%	61.83%
STRA-Net	RGB+ Flow	92.9%	73.66%

4.3.4. Visualization

In order to visually illustrate the effectiveness of the proposed method, the Grad-CAM [18] method is used to visualize the features concerned by the last layer of convolution of the feature extraction part, and the highlighted areas of the heatmap reflect the areas of focus of the model, and the model determines the results of action video classification through these key areas. In order to locate the position of the heatmap in the original image, the heatmap is superimposed with the original image in a certain proportion. As shown in Fig. 8, it can be clearly seen from the figure that for the jump rope action, the action recognition method based on temporal and spatial residual attention can capture the trajectory of human jump when jumping rope.

In order to more precisely analyze the recognition effect of this method and the baseline network on the dataset HMDB51 single class, the single class classification accuracy of the two methods is shown in Fig. 9. As shown in the figure, the horizontal ordinate represents 51 action videos, the vertical ordinate represents the recognition accuracy, the blue column represents the action recognition accuracy of each class on the HMDB51 dataset based on spatio-temporal residual attention two-stream network, and the orange column represents the action recognition accuracy of each class on the HMDB51 dataset based on the baseline network. According to statistics, in 51 kinds of action video, 43 kinds of action video recognition accuracy are higher than or equal to the recognition accuracy of the baseline network method. It can be seen that our model has good recognition performance.



Fig. 8 Feature area heatmap

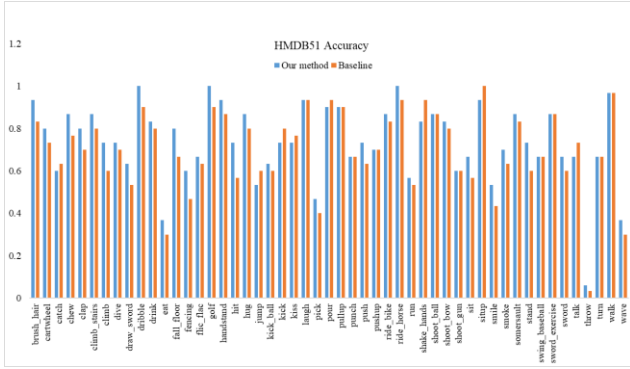


Fig. 9 Recognition rate for each category on the HMDB51 dataset

Table 3. Comparisons with the State-of-the-arts

Method	UCF101(%)	HMDB51(%)
LRCN [17]	87.9	
T3D-Transfer [30]	91.7	61.1
Two-Stream [9]	88.0	59.4
C3D [18]	85.2	59.1
Multi-branch attention [30]	90.6	61.7
Hierarchical Attention [31]	92.7	64.3
JNA + Warped JNA [32]	91.5	68.8
Video LSTM [33]	88.9	56.4
LTC [34]	91.7	64.8
LSTM spatial transformer [35]	92.8	67.1
D-Res-LSTM [36]	92.3	
TS FCAN-comp [37]	92.0	66.7
Our method	92.9	73.66

5. Conclusion

Aiming at the problem that the two-stream methods ignore the difference of different channel information of features, and cannot distinguish the redundant frames of video, background and other spatio-temporal information, resulting in weak overall feature expression ability and low recognition rate, we propose an action recognition method for two-stream spatio-temporal residual attention network. By adding a convolution block attention module (CBAM) to the network, the model not only focuses on spatial features, but also on channel features, which further improves the performance of the model. In this paper, the UCF101 and HMDB51 datasets are verified, and the experimental results are compared with those of other advanced algorithms. All the results show that

4.4. Comparisons with the State-of-the-arts

In this section, the best performance obtained from the experiment is compared with the existing advanced algorithms, and the recognition accuracy of UCF101 and HMDB51 is listed respectively. The final results are shown in Table 3. From the results of the table, the method proposed in this paper is superior to other advanced algorithms. Compared with the HMDB51 dataset, the accuracy improvement on the UCF101 dataset is not too high. It may be because the HMDB51 dataset is a video intercepted from various film websites. There is a lot of interference information in the video, such as occlusion, viewpoint change, background difference and confusion. This paper introduces the attention mechanism to effectively capture and focus on key information such as key areas and important channels of the video, reduce the interference of background information, and improve the accuracy of action recognition. Specifically, compared with JNA + Warped JNA, the proposed method improves the accuracy by 1.4 % and 4.86 % on UCF101 and HMDB51 datasets, respectively. Compared with LSTM spatial transformer method, the proposed method improves 0.1 % and 6.56 % on these two datasets respectively. Compared with the 3D convolution model method LTC, the accuracy of the proposed method on UCF101 and HMDB51 datasets increased by 1.2 % and 8.86 %, respectively.

the model improves the accuracy of action recognition to a certain extent. In the next work, consider how to improve recognition accuracy in video datasets with higher complexity and more temporal information, such as the Something - Something V1 / V2 dataset.

References

- [1] Fan Z, Zhao X, Lin T, et al. Attention-based multiview re-observation fusion network for skeletal action recognition [J]. IEEE Transactions on Multimedia. Vol. 21(2019) No. 2, p. 363-374.
- [2] LIAO Z, HU H, LIU Y. Action recognition with multiple relative descriptors of trajectories[J]. Neural Processing Letters, 2020, 51(1):287-302.

- [3] SHI L,ZHANG Y,CHENG J,et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C] // Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach:IEEE,2019:12018-12027.
- [4] LI T, FAN L, ZHAO M, et al. Making the invisible visible:action recognition through walls and occlusions [C]// Proceedings of 2019 IEEE International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019:872-881.
- [5] Wang H , Klser A , Schmid C , et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International Journal of Computer Vision, 2013(1):103.
- [6] Laptev I, Lindeberg T. On space-time interest points[J]. International Journal of Computer Vision, 2005, 64(2-3): 107-124.
- [7] Chaudhry R , Ravichandran A , Hager G , et al. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2009.
- [8] Bobick A F , Davis J W . The recognition of human movement using temporal templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(3):257-267.
- [9] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27.
- [10] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]. European Conference on Computer Vision. Springer, Cham, 2016: 20-36.
- [11] Fischer P , Dosovitskiy A , Ilg E , et al. FlowNet: Learning Optical Flow with Convolutional Networks[J]. IEEE, 2016.
- [12] Hochreiter S , Schmidhuber J . Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.
- [13] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4694-4702.
- [14] Ballas N, Yao L, Pal C, et al. Delving deeper into convolutional networks for learning video representations[J]. arXiv preprint arXiv:1511.06432, 2015.
- [15] Gammulle H, Denman S, Sridharan S, et al. Two stream lstm: A deep fusion framework for human action recognition[C]. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017: 177-186.
- [16] Li Z, Gavriluk K, Gavves E, et al. Videolstm convolves, attends and flows for action recognition[J]. Computer Vision and Image Understanding, 2018, 166: 41-50.
- [17] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2625-2634.
- [18] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.
- [19] Du T , Ray J , Shou Z , et al. ConvNet Architecture Search for Spatiotemporal Feature Learning[J]. 2017.
- [20] Carreira J, Zisserman A. Quo Vadis, action recognition? A new model and the kinetics dataset[C]. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.
- [21] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 5533-5541.
- [22] Du T , Wang H , Torresani L , et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition[J].
- [23] Sun L , Jia K , Yeung D Y , et al. Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks: IEEE, 10.1109/ICCV.2015.522[P]. 2015.
- [24] Zhou Y , Sun X , Zha Z J , et al. MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.
- [25] Hu Jie, Shen Li, Albanie S, et al. Squeeze-and-excitation networks [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.
- [26] ZHANG J, XIE Y, XIA Y, et al. Attention residual learning for skin lesion classification[J]. IEEE Transactions on Medical Imaging, 2019, 38(9): 2092-2103.
- [27] MAX J, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[C]//Proceedings of ECCV'15. Berlin, Germany: Springer, 2015: 2017-2025.
- [28] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, pages 4489–4497, 2015.
- [29] Diba A, Fayyaz M, Sharma V, et al. Temporal 3D Convnets: New Architecture and Transfer Learning for Video Classification[J]. arXiv preprint arXiv: 1711.08200, 2017.
- [30] Jialin Wu, Gu Wang, Wukui Yang, Xiangyang Ji. Action Recognition with Joint Attention on Multi-Level Deep Features.[J]. CoRR, 2016, abs/1607.02556.
- [31] Wang Y , Wang S , Tang J , et al. Hierarchical Attention Network for Action Recognition in Videos[J]. 2016.
- [32] Zang, J. et al. Attention-based Temporal Weighted Convolutional Neural Network for Action Recognition. (2018).
- [33] Li, Z., Gavriluk, K., Gavves, E., Jain, M. & Snoek, C. G. M. VideoLSTM convolves, attends and flows for action recognition. Computer Vision and Image Understanding 166, 41–50 (2018).
- [34] Varol G, Laptev I, Schmid C. Long-Term Temporal Convolutions for Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1510-1517.
- [35] Ge H W, Yan Z H, Yu W H, et al. An Attention Mechanism Based Convolutional LSTM Network for Video Action Recognition[J]. Multimedia Tools and Applications, 2019, 78(14): 1-24.
- [36] Song S, Liu J, Li Y, Guo Z (2020) Modality compensation network: Cross-modal adaptation for action recognition. IEEE Trans Image Process 29:3957–3969.
- [37] Tran A, Cheong L-F (2017) Two-stream fow-guided convolutional attention networks for action recognition. In: Proceedings of the IEEE international conference on computer vision. pp 3110–3119.