Journal of Computers Vol. 33 No. 1, February 2022, pp. 21-30
doi: 10.53106/199115992022023301003

Cascade Attention-based Spatial-temporal Convolutional Neural Network for Motion Image Posture Recognition

Shuqi Zhang*

College of Physical, Zhengzhou University of Science and Technology, Zhengzhou 450064, China
352720214@qq.com

Received 27 September 2021; Revised 15 October 2021; Accepted 27 October 2021

Abstract. The traditional motion posture recognition methods cannot capture the temporal relationship in a video sequence, which leads to the problem that the recognition effect of time-dependent behaviors is not ideal. Therefore, this paper proposes a cascade attention-based spatial-temporal convolutional neural network for motion posture recognition. Firstly, the convolutional neural network is used to model the time sequence relationship in the video, so as to capture the spatial-temporal information in the video efficiently. At the same time, the cascade attention mechanism is used to improve the low learning ability of spatial features caused by channel information moving on the time axis. Meanwhile, a new spatial-temporal network structure is constructed, which includes the spatial-temporal appearance information flow and spatial-temporal motion information flow. Finally, the weighted average method is used to fuse the two spatial-temporal networks to obtain the final recognition result. Experiments are conducted on UCF101 and HMDB51 datasets, respectively, and the recognition accuracy is 96.8% and 79.6%. Experiment results show that compared with the state-of-the-art network methods, the recognition accuracy with the proposed method has better effect and robustness.

Keywords: motion posture recognition, cascade attention, spatial-temporal convolutional neural network, weighted average

1 Introduction

As a key part of video understanding, motion posture recognition has always been a research hotspot in the field of computer vision. It has high application value in the fields of video surveillance, virtual reality, intelligent human-machine interface and social video recommendation [1-3]. Because of the complex background, the appearance difference of objects and the similarity of different types of motions in real scenes, motion recognition is still a challenging topic [4].

Video motion recognition can be mainly divided into manual design-based features [5] and deep learning-based methods [6]. The latter shows better performance, in which the dual-stream convolutional network method [7-9] can effectively extract the appearance information and motion information from the video, and achieve a better recognition effect in the motion recognition task. However, it is still difficult to utilize the spatial-temporal information in video effectively. Therefore, researchers put forward a variety of improvement methods. In terms of network input, Bilen et al. [10] compressed the video sequence into a dynamic image on the condition that the order information was retained, and used it as the input of the deep network to extract the time sequence information in the video. However, the generation of dynamic image brought complicated calculation processes. In terms of network structure, Feichtenhofer et al. [11] used the residual network to build a dual-flow network model, and proposed to add a short connection between two convolutional streams to enhance the information interaction between the dual-flow networks. For network fusion, reference [12] fused two networks in the middle of the hidden layer to make the network learn the relationship between temporal domain features and spatial domain features. And a variety of fusion methods were proposed. Although the improvements in the above three aspects can make the double-stream network better use the spatial-temporal information in the video and improve the accuracy of behavior recognition, there is still a problem that the time-sequence relationship in the video sequence cannot be captured. In addition, the 3D convolutional neural network [13] also had a good performance in human behavior recognition, but its parameters and calculation amount would be greatly increased. Therefore, Lin et al. [14] proposed a Temporal Shift Module (TSM), which used two-dimensional convolutional neural network to extract Temporal sequence information in videos, but it reduced the learning ability of spatial features of the network.

Therefore, this paper proposes a cascade attention-based spatial-temporal convolutional neural network for motion posture recognition. This new method shows better effectiveness on motion posture recognition through

* Corresponding Author

rich experiments.

2 Proposed Motion Posture Recognition

2.1 The Overall Framework

The overall structure of the proposed cascade attention-based spatial-temporal convolutional neural network (CA-CNN) in this paper is shown in Fig. 1. It is divided into three parts: video segment random sampling, cascade attention spatial-temporal network and spatial-temporal network fusion [15]. Firstly, the input video is randomly sampled in segments, and then the sampled RGB video frames and a group of optical flow images are fed into the new spatial-temporal network. And the initial class scores of the video in the spatial-temporal appearance information flow (STAIF) and the spatial-temporal motion information flow (STMIF) are obtained. Finally, the weighted average method is adopted to fuse the initial class scores, and the final recognition results are obtained through Softmax.

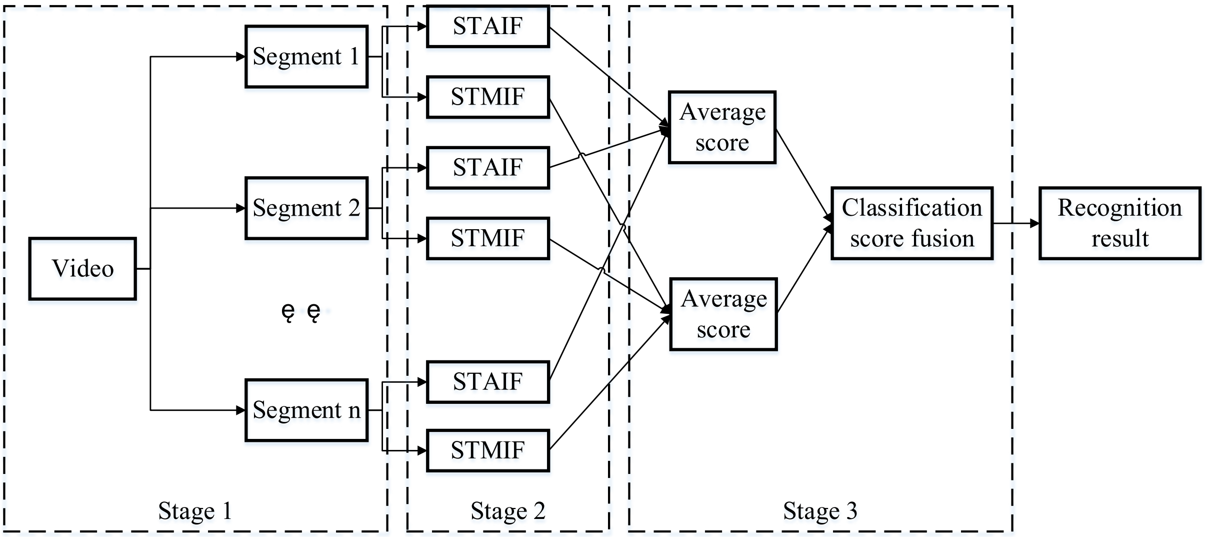

Fig. 1. Overall structure of the proposed algorithm in this paper

2.2 Video Segment Random Sampling and Network Fusion

The existing double-stream network methods had achieved good results in the recognition of short-time behavior, however, only single RGB video frames (spatial flow) and stacked optical flow images (time flow) can be used to learn the appearance features and motion features. Therefore, in the process of recognizing motions with a long time span, some important information will be lost. As a result, the learned features cannot accurately represent the whole motion, and it is difficult to accurately recognize long-term motions. For this reason, video segment random sampling strategy is adopted to realize the effective learning of the whole motion video, and sparse sampling method reduces the redundant information in the video. Specifically, the input video is divided into K segments $\{S_1, S_2, \cdots, S_K\}$ with equal time length. The sequence of fragments is then modeled as follows:

$$H_\alpha = g(P_\alpha(T_{\alpha 1};W_\alpha), P_\alpha(T_{\alpha 2};W_\alpha), \cdots, P_\alpha(T_{\alpha k};W_\alpha)) . \quad \textbf{(1)}$$

$$H_\beta = g(P_\beta(T_{\beta 1};W_\beta), P_\beta(T_{\beta 2};W_\beta), \cdots, P_\beta(T_{\beta k};W_\beta)) . \quad \textbf{(2)}$$

$$CA - CNN(T_1, T_2, \cdots, T_k) = \max(\delta(\lambda H_\alpha + \mu H_\beta)) . \quad \textbf{(3)}$$

Wherein, subscripts α and β are used to distinguish spatial-temporal appearance information flow and spatial-temporal motion information flow. T_i represents the sequence of segments randomly sampled from the corresponding video segment $S_i(i=1,2,\cdots,K)$. $T_{\alpha i}$ is video frame. $T_{\beta i}$ is an optical flow image. P_α and P_β are functions that calculate the score of $T_{\alpha i}$ and $T_{\beta i}$ belonging to each category. W_α and W_β are the network parameters of spatial-temporal appearance information flow and spatial-temporal motion information flow. g is a fusion function that averages all the scores for which T_i belongs to the same category. H_α and H_β are the category scores of spatial-temporal appearance information flow and spatial-temporal motion information flow respectively. λ and μ are the proportionality coefficients of double-flow fusion. δ is the Softmax function, which is used to predict the probability that the entire video falls into each behavior category. The category with the highest probability is judged as the behavior to which the video belongs.

In addition, network parameters are shared between K segments (K=3 in this paper). Combining with the standard cross entropy loss, the final loss function is:

$$L(y,H)=-\sum_{i=1}^{C}y_i(H_i-\log\sum_{j=1}^{C}\exp(H_j)) \cdot \qquad \textbf{(4)}$$

Where C is the number of motion categories. $H= g(P(T_1;\ W),\ P(T_2;\ W),\ ...,\ P(T_K;\ W))$. y_i is the true label for i-th motion. H_i is the classification score of i-th class motion. The proposed dual-stream network learning is a non-end-to-end process, that is, the two networks are trained and tested separately, then they are fused.

Combining with the standard back propagation algorithm, multiple fragments are used to jointly optimize the network parameter W. In the process of back propagation, the gradient of network parameter W related to the loss value L can be expressed as:

$$\frac{\partial L(y,H)}{\partial W}=\frac{\partial L}{\partial H}\sum_{k=1}^{K}\frac{\partial g}{\partial P(T_k)}\frac{\partial P(T_k)}{\partial W} \ . \qquad \textbf{(5)}$$

2.3 Cascade-Attention Spatial-Temporal CNN

A. Temporal Shift Module

The rapid growth of video streaming has brought great challenges to the video understanding. The processing of massive video requires low computational cost and high precision. At present, 3D convolution has good performance in extracting spatial-temporal features, but it is expensive to deploy due to intensive computation. So, the temporal shift module (TSM) was proposed with high efficiency and high performance, which could realize the performance similar to three-dimensional convolution with the complexity of two-dimensional convolution.

The TSM decouples the convolution process into two steps: data shift and multiply-accurate respectively. Wherein, the convolution operation $Y=Conv(W,X)$ can be expressed as $Y=\omega_1X_{i\text{-}1}+\omega_2X_i+\omega_3X_{i+1}$. The convolution weight $W=(\omega_1,\omega_2,\omega_3)$. The input X is a one-dimensional vector with indefinite length. As shown in Fig. 2, data movement of -1, 0, +1 is carried out on some channels in the time dimension, so that the information from adjacent frames is mixed with the information of the current frame after movement, so as to realize the timing modeling of the video. It can be formally expressed as $X_i^{-1}=X_{i-1}$, $X_i^0=X_i$, $X_i^{+1}=X_{i+1}$. Then they multiply the weights $(\omega_1,\omega_2,\omega_3)$ respectively, and obtain $Y=\omega_1X^{-1}+\omega_2X^0+\omega_3X^{+1}$. The first “shift” can be performed without any multiplication. However, the calculation cost of the second step is relatively high. In order not to add additional parameters and calculation cost, TSM combines the "multiply-accurate" into the convolutional neural network, so no additional calculation amount will be added.

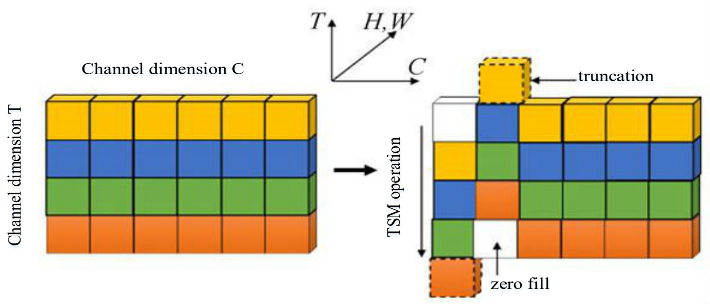

Fig. 2. TSM operation

The feature mapping in a video model can be expressed as: $A \in R^{N \times C \times T \times H \times W}$. Where N is the batch size, C is the number of channels, T is the time dimension. H and W are the spatial resolutions. Two-dimensional convolution independently works on the time dimension T, and the information of each channel exists independently, so it does not have the ability of time series modeling. For this reason, TSM module is introduced in this paper. By moving part of the channel forward and backward along the time dimension T, the channel information from the image sequence of adjacent fragments (i.e. in the spatial-temporal appearance information stream, it is RGB video frame. In the spatial-temporal motion information flow, it is an optical flow image) is mixed to realize the modeling of the video timing relationship. Thus, the motion appearance features and motion features containing time series information are extracted.

B. Cascade Attention Module

In order to solve the problem of the decline of spatial feature learning ability caused by time shift, this paper introduces the cascade attention module to make the network learn the key local details by applying channel attention and spatial attention in the channel and spatial dimensions, so as to enhance the feature learning and expression ability of the network.

The structure of cascade attention is shown in Fig. 3. Given an intermediate feature map $F \in R^{C \times H \times W}$ as input, the one-dimensional channel attention map $M_c \in R^{C \times 1 \times 1}$ and the two-dimensional space attention map $M_s \in R^{1 \times H \times W}$ are input successively. The calculation process of the whole attention can be summarized as:

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_c(F') \otimes F' \end{aligned} \tag{6}$$

Where $\otimes$ represents element multiplication. During multiplication process, the attention value is broadcast accordingly. Channel attention values are broadcast along spatial dimensions. Spatial attention values are broadcast along channel dimensions. F'' is the final output.

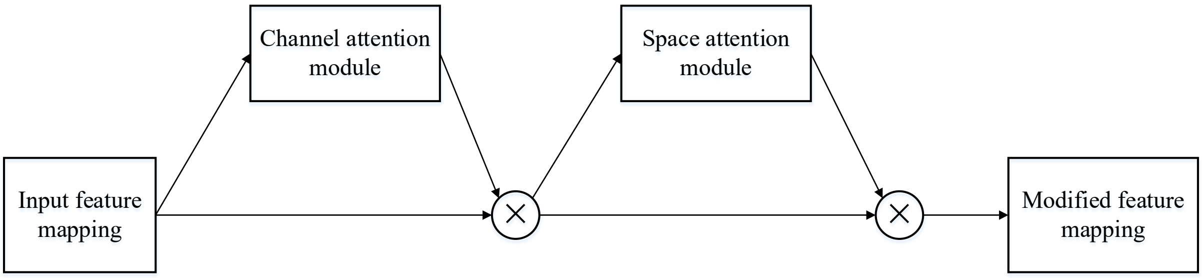

Fig. 3. Cascade attention module

The channel attention module uses global maximum sum average pooling to aggregate the spatial information of feature maps to generate two different spatial context descriptions: F_{avg}^{c} and $F_{\max}^{c}$. Then, a shared network composed of multi-layer perceptron (MLP) is used to calculate the two different spatial context descriptions to obtain channel attention feature map $M_c \in R^{C\times1\times1}$. The specific calculation process is shown in equation (7).

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)\ + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F_{avg}^{c})\ + W_1(W_0(F_{\max}^{c}))) \end{aligned} \quad . \tag{7}$$

Where σ denotes the Sigmoid function, $W_0 \in R^{C/r\times C}$, $W_1 \in R^{C/r\times C}$, r is the reduction ratio. The weights W_0 and W_1 of the MLP are shared for both inputs, and the W_0 is followed by the ReLU activation function.

The spatial attention module takes the feature mapping output by the channel attention module as the input of the module, and uses global maximum sum average pooling in the channel dimension to get two different feature descriptions: $F_{avg}^{S} \in R^{1\times H\times W}$ and $F_{\max}^{S} \in R^{1\times H\times W}$. Then the two feature descriptions are matched in a cascade way. The convolution operation is used to generate the spatial attention feature map $M_s(F) \in R^{1\times H\times W}$. The calculation process of spatial attention is as follows:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}([F_{avg}^{s}; F_{\max}^{s}])) \end{aligned} \quad . \tag{8}$$

Where σ denotes the Sigmoid function. $f^{7\times7}$ is a convolution operation with a convolution kernel size of 7×7.

C. Network structure

The specific network structure of spatial-temporal appearance information flow and spatial-temporal motion information flow is shown in Fig. 4. Both streams are based on the Resnet50 network [16]. TSM is added into the residual block by residual shift. The spatial-temporal appearance information flow of cascade attention is placed in the residual block. The position after the spatial-temporal motion information flow is introduced to the last convolution layer. The above designs have the best effective in experiments. Because the method of residual displacement can make the network capture time series information better and alleviate the degradation of spatial feature learning ability caused by information capture to a certain extent. Due to the spatial-temporal appearance information flow, the RGB video frame input contains complex scene information. Adding cascade attention to the residual block for feature calibration can make the network learn more accurate spatial appearance features. In the spatial-temporal motion information flow, since the input optical flow image only contains motion information, the original network can be used to complete feature extraction, and the fine-tuning of high-level features by using cascade attention can enable the network to obtain more accurate feature expression.

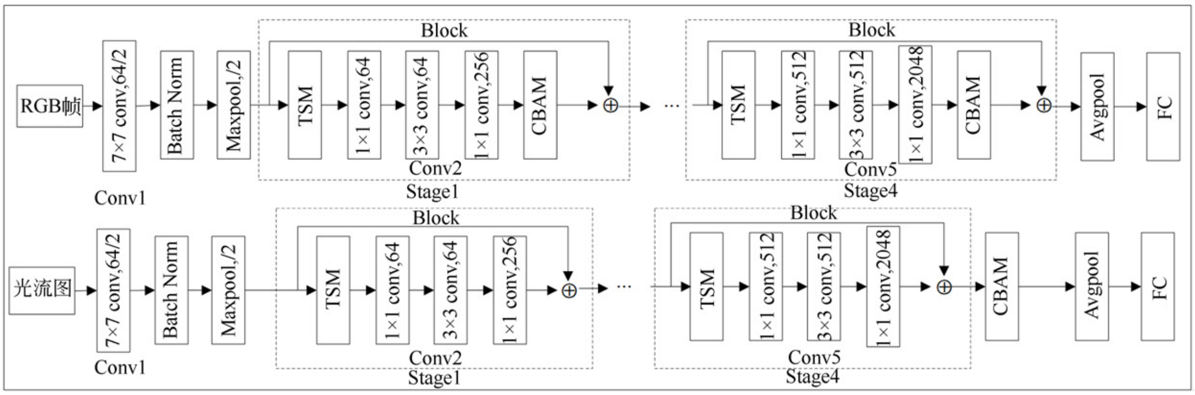

Fig. 4. Modified spatial-temporal convolution neural network

3 Experiments and Analysis

3.1 Data Set

The proposed algorithm is evaluated on the UCF101 [17] and HMDB51 [18] data sets, and the experimental results are compared with the current mainstream motion recognition algorithms to verify the effectiveness. The UCF101 is mainly realistic videos from YouTube, which are affected by camera motion, complex scenes, lighting changes, occlusion, video blur and other factors. It includes 101 motion categories and 13320 videos. According to the types of motions in the video, they can be divided into five categories: person-to-person interaction, person-to-object interaction, body movement, musical instrument playing and sports. HMDB51 is mainly composed of movie clips, including 6766 videos, a total of 51 motion categories. Each category contains at least 100 videos, including individual behavior, facial expression and object manipulation behavior, person-to-person interaction, person-to-object interaction, etc. Three groups of training sets and test sets are divided according to the official methods provided by the two data sets. The recognition accuracy obtained from the test sets of the three partitioning methods is averaged as the final recognition result in this paper.

3.2 Experiment Settings

The experiment environment is: Pytorch1.4.0, CUDA10.0, CUDNN 7.6.5, Ubuntu18.04 system. The computer is configured as Intel Xeon(R) Silver 4112, CPU 2.6GHz, NVIDA GeForce1060 TI graphics. Small batch stochastic gradient descent method is used, and the momentum is 0.9. According to the memory size and GPU utilization of the computer, the batch size is set to 8, the initial learning rate is 0.001. The training time is 25 epochs. The attenuation rate is 0.1 for every 10 epochs. The optical flow image is calculated by using TVL1 in OpenCV library combining denseflow tool library and GPU.

Due to the small size of the data set used in the experiment, in order to avoid over-fitting in the training process, the network adopts weight initialization pre-trained on the ImageNet+Kinetics database. And it uses corner clipping and multi-scale clipping methods to make data augmentation. The randomly sampled images with 340×256pixels are cropped. In corner cropping, the image is cropped from the center and four diagonals to get 224×224pixels. In multi-scale cropping, two values are randomly selected from {168,192,224,256} in the center and four diagonal angles as the width and height of the image for clipping, and then the image is adjusted to 224×224pixels. In addition, the mean and variance parameters in BN of other convolutional layers except the first layer are fixed. During the testing, each video segment is sampled twice. Eight groups of RGB frames or optical flow images are sampled each time. After the sample image is scaled, the left/right corners and the center are cropped, and a full resolution image with a shorter side of 256×256 pixels is used for testing.

3.3 Ablation Experiment

In order to verify the relative importance of the time shift module and the cascade attention module in the improved spatial-temporal convolution neural network, the following ablation experiments are carried out.

In order to verify the effectiveness of the time shift module (TSM), the recognition accuracy of the proposed algorithm in this paper on the UCF101 and HMDB51 datasets are compared. The segment number is set as 3. According to the original parameter setting of the time shift module, the reciprocal of the shift ratio is set to 8. The experimental results are shown in Table 1.

Table 1. Comparison of accuracy after adding TSM/%

Backbone network	UCF101		HMD51	
	STAIF	STMIF	STAIF	STMIF
ResNet50	89.3	89.8	59.5	71.9
ResNet50+TSM	93.3	94.5	67.2	74.3

According to the results in Table 1, the recognition accuracy of spatial-temporal appearance information flow and spatial-temporal motion information flow are improved by 4.0% and 4.7% in UCF101 and 7.7% and 2.4% in HMDB51 respectively after adding TSM.

In the experimental results, it is found that the recognition accuracy of some behaviors is increased, while some behaviors is decreased. For the convenience analysis, ten behaviors with the largest increase and decrease values in accuracy are selected as shown in Table 2.

Table 2. Change rate of top 10 motions after adding TSM

UCF101		HMD51	
Motions	Δ Accuracy	Motions	Δ Accuracy
Jump rope	52.6	Clap	30.0
Front Crawl	48.6	Wave	23.3
Jumping jack	45.9	Climb	23.3
Nun chucks	45.7	Ran	23.3
High jump	40.5	Push	20.0
Baseball pitch	-25.5	Kick ball	-16.7
Brushing teeth	-25.0	Stand	-10.0
Shaving beard	-18.6	Ride horse	-10.0
Tennis swing	-14.3	Sword	-6.7
Ski jet	-14.2	Kiss	-6.7

The above phenomenon may be caused by the fact that the convolutional neural network can extract the spatial-temporal information in the video after adding the time shift module. However, the convolutional neural network loses some learning ability of spatial features while capturing the time sequence information, which leads to the degradation of the behavior recognition effect that is highly dependent on the spatial scene information. For example, in Baseball Pitch and Tennis Swing, better recognition results can be obtained by relying on the Baseball field and Tennis court in the scene, while the recognition accuracy has decreased after adding TSM. However, Jump Rope, Jumping Jack and High Jump are highly dependent on temporal information, and the accuracy improvement proves that it is feasible to extract spatial-temporal information from behavior video by introducing TSM to convolutional neural network.

In order to verify the effectiveness of cascade attention (CA), the comparison is conducted with ResNet50 and ResNet50 +TSM in the same experimental setting. The results are shown in Table 3. The best recognition results are obtained when (c) and (a) are combined.

Table 3. Recognition accuracy with different networks

Backbone network	UCF101		HMD51	
	STAIF	STMIF	STAIF	STMIF
ResNet50	88.3	89.8	58.5	70.9
ResNet50+CA(a)	88.2	90.1	59.3	71.3
ResNet50+CA(b)	88.5	89.8	59.9	70.3
ResNet50+CA(c)	88.5	89.3	60.4	69.4
ResNet50+TSM	92.3	93.5	66.2	73.3
ResNet50+TSM+CA(a)	92.5	93.8	66.6	73.9
ResNet50+TSM+CA(b)	92.6	93.7	66.6	73.5
ResNet50+TSM+CA(c)	92.9	93.4	66.8	73.3

Note: where (a), (b) and (c) represent the CA added after the last convolution layer, after the first convolution layer and the last convolution layer, and after the last convolution layer in Resnet50, respectively.

Table 4 shows the ten motions with the greatest improvement in accuracy after adding the CA. For Baseball Pitch, Shaving Beard, Tennis Swing, Kick Ball and Ride Horse, the recognition accuracy has been greatly improved, which solves the problem of decreased learning ability of spatial features brought by the TSM to a certain extent. For Laugh, Drink, Throw, Catch and other facial or hand motions, and the similar behaviors such as Climb Stairs, Climb, Throw, Catch, the recognition is more accurate, which proves that the method combined with the CA can enhance the feature learning and expression ability of the network. It can enable the network to learn more detailed behavior features, and improve the recognition ability of similar behaviors.

Table 4. Change rate of top 10 motions after adding CA

UCF101		HMD51	
Motions	Δ Accuracy	Motions	Δ Accuracy
Hammering	30.3	Throw	30.0
Wall pushups	25.7	Laugh	16.7
Baseball pitch	23.3	Climb stairs	13.3
Shaving beard	23.3	Fall floor	13.3
Tennis swing	22.4	Drink	13.3
Skate boarding	18.7	Kick ball	13.3
Rafting	17.8	Catch	10.0
Tai chi	17.8	Ride horse	10.0
skiing	15.0	Climb	6.7
Boxing punching-Bag	14.3	Dribble	6.7

3.4 Improved Spatial-temporal Network Fusion Experiment

Finally, the classification scores of spatial-temporal appearance information flow and spatial-temporal motion information flow are fused by weighted average. We search the optimal proportion coefficient by experiment. Due to the higher accuracy in the spatial-temporal motion information flow, it is attempted to give it a larger weight coefficient. As can be seen from Table 5, when the fusion ratio is 1:1.8, the accuracy does not increase, and the highest average accuracy is obtained at this time.

Table 5. Recognition accuracy with different fusion ratios/%

Fusion rations	UCF101				HMDB51			
	K=1	K=2	K=3	Average	K=1	K=2	K=3	Average
1:1	95.7	96.5	95.9	96.0	76.4	76.8	76.3	76.5
1:1.2	95.8	96.6	96.1	96.1	77.2	77.2	76.9	77.1
1:1.4	96.0	96.6	96.2	96.2	77.7	77.2	77.7	77.5
1:1.6	96.1	96.6	96.2	96.2	78.0	77.4	78.0	77.8
1:1.8	96.2	96.5	96.5	96.4	78.2	77.3	77.9	77.8
1:2.0	96.1	96.4	96.6	96.3	77.9	76.9	77.8	77.5
1:2.2	95.9	96.3	96.5	96.2	77.8	76.7	77.7	77.4
1:2.4	96.1	96.3	96.4	96.2	77.7	76.8	77.7	77.4

In order to reflect the advantages of the proposed algorithm in recognition accuracy, UCF101 and HMDB51 datasets are selected for experiments. The proposed algorithm is compared with the existing mainstream motion recognition algorithms including DMHI [19], DRT [20], KRF [21], REA [22]. The recognition accuracy of each algorithm is shown in Table 6.

It can be seen from Table 6 that the recognition accuracy of the proposed algorithm has certain advantages over the existing motion recognition algorithms. The reason is that the proposed cascade attention spatial-temporal network in this paper can effectively utilize the temporal relationship information and spatial information in the video, and improve the ability to identify the time-series dependent motions. In addition, the ability to learn local details in the network space is enhanced, and similar motions can be better recognized, thus improving the recognition accuracy.

Table 6. Accuracy comparison with different algorithms/%

Method	UCF101	HMDB51
DMHI	79.4	66.7
DRT	88.6	69.5
KRF	92.7	73.2
REA	94.1	77.8
Proposed	96.8	79.6

4 Conclusions

This paper proposes a cascade attention-based spatial-temporal convolutional neural network for motion posture recognition. By combining the time shift and attention mechanism, a new spatial-temporal network structure is constructed, which includes the spatial-temporal appearance information flow and spatial-temporal motion information flow. It realizes the effective extraction of time and space features in the video. At the same time, the cascade attention module is used to emphasize the key detail features in the channel and space, which enhances the feature expression ability of the network, and thus improves the ability to identify the time-series dependent motions and similar motions. Experimental results show that the recognition accuracy of the proposed algorithm on motion recognition datasets UCF101 and HMDB51 is 96.8% and 79.6% respectively, which is higher than that of the existing algorithms. In order to further improve the recognition performance, it can also be improved from the perspective of building an end-to-end spatial-temporal network.

5 Acknowledgement

The authors thank for the anonymous review.

References

[1] J.-S. A, S.-L. Yin, A New Feature Fusion Network for Student Behavior Recognition in Education, Journal of Applied Science and Engineering 24(2)(2021) 133-140.

[2] S.-W. Ma, L.-N. Liu, Q. Fu, J.-R. Wen, Using PHOG fusion features and multi-class Adaboost classifier for human behavior recognition, Guangxue Jingmi Gongcheng/Optics and Precision Engineering 26(11)(2018) 2827-2837.

[3] J. Candamo, M. Shreve, D.-B. Goldgof, D.-B. Sapper, R.Kasturi, Understanding Transit Scenes: A Survey on Human Behavior-Recognition Algorithms, IEEE Transactions on Intelligent Transportation Systems 11(1)(2010) 206-224.

[4] Y. Wang B.-Y. Wang, Y.-Z. Yu, Q.-H. Dai, Z.-W. Tu, Action-Gons: Action Recognition with a Discriminative Dictionary of Structured Elements with Varying Granularity, In: Computer Vision -- ACCV 2014. (2015). Lecture Notes in Computer Science, vol 9007. Springer, Cham.

[5] Y. Zhu, J.-K. Zhao, Y.-N. Wang, B.-B. Zheng, A Review of Human Action Recognition Based on Deep Learning, Acta Automatica Sinica 42(6)(2016) 848-857.

[6] D. Zheng, H. Li, S.-L. Yin, Action Recognition Based on the Modified Two-stream CNN, International Journal of Mathematical Sciences and Computing (IJMSC) 6(6)(2020) 15-23.

[7] S.-L. Yin, J. Liu, L. Teng, A Sequential Cipher Algorithm Based on Feedback Discrete Hopfield Neural Network and Logistic Chaotic Sequence, International Journal of Network Security 22(5)(2020) 869-873.

[8] L. Wang, Y. Xiong, W. Zhe, Q. Yu, L.-V. Gool, Temporal Segment Networks for Action Recognition in Videos, IEEE Transactions on Pattern Analysis and Machine Intelligence 41(11)(2019) 2740-2755.

[9] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, Action Recognition with Dynamic Image Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 40(12)(2018) 2799-2813.

[10]C. Feichtenhofer, A. Pinz, R.-P. Wildes, Spatiotemporal Multiplier Networks for Video Action Recognition, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11]C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional Two-Stream Network Fusion for Video Action Recognition, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[12]D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, in: Proc. 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[13]J. Lin, C. Gan, S. Han, TSM: Temporal Shift Module for Efficient Video Understanding, in: Proc. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[14]S. Woo, J. Park, J.-Y. Lee, I.-S. Kweon, CBAM: Convolutional Block Attention Module, ECCV 2018. Lecture Notes in Computer Science, vol 11211. (2018)

[15]N. Shibata, M. Tanito, K. Mitsuhashi, Y. Fujino, M. Matsuura, H. Murata, R. Asaoka, Development of a deep residual learning algorithm to screen for glaucoma from fundus photography, Scientific Reports 8(1)(2018).

[16]K. Soomro, A.-R. Zamir, M. Shah, UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, Computer Science, (2012), arXiv:1212.0402.

[17]H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: A large video database for human motion recognition, in: Proc. 2011 International Conference on Computer Vision, 2011.

[18]M. Murakami, J.-K. Tan, H. Kim, S. Ishikawa, Human motion recognition using directional motion history images, in: Proceedings of International Conference on Artificial Life and Robotics 25, 2020.

[19]H.-L. Yang, M.-Z. Huang, Z.-Q. Cai, Research on Human Motion Recognition Based on Data Redundancy Technology, Complexity (4)(2021) 1-6.

[20]Y.-J. Zhou, C.-A. Di, Human motion recognition based on Kalman random Forest algorithm and 3D multimedia, Multimedia Tools and Applications 79(2)(2020).

[21]Y. Wang, F. Feng, Reliability Enhancement Algorithm of Human Motion Recognition Based on Knowledge Graph, International Journal of Distributed Systems and Technologies (IJDST) 12(2021).