https://doi.org/10.31449/inf.v48i19.6188

Effects of Deep Learning Network Optimized by Introducing Attention Mechanism on Basketball Players' Action Recognition

Xianyou Yan
Department of P.E., Henan University of Animal Husbandry and Economy, Zhengzhou 450046, China
E-mail: 80448@hnuahe.edu.cn

Keywords: two-stream, attention mechanism, spatio-temporal network, deep learning, basketball action

Received: May 14, 2024

Video, as an important carrier of big data storage, can help people to achieve behavioral analysis as well as localization. This study presents a basic network constructed and optimized based on the two-stream fusion algorithm, and introduces an attention mechanism to enhance its performance. First, the short-range action timing feature is extracted by optical flow feature, and the spatial feature is extracted by RGB frame, and the two features are combined by double-flow method to improve the ability of capturing spatio-temporal features. Multi-head self-attention module and spatial attention mechanism are introduced to improve the capturing effect of temporal and spatial features. Experiments were carried out on HMDB51 and Charades-STA datasets. The results showed that the recognition accuracy of MHSSA module in time series feature extraction is superior to the traditional Softmax mechanism, with an average recognition accuracy of more than 85%. In the HMDB51 dataset, the accuracy of basketball action recognition was 95.4%, and the recognition speed was 20 frames per second. Compared with the GTN and DGCN models, the model proposed in study had the best overall performance in the Charades-STA and ActivityNet-Captions datasets, which were 7.86% and 6.92% higher, respectively. In addition, the recognition speed of this method was 25% faster than other models, and the average recognition time was reduced by 47.27%. Consequently, the proposed optimized spatio-temporal neural network based on a two-stream architecture is capable of performing basketball sports action recognition with greater speed and accuracy

Povzetek: Raziskava predstavlja optimizirano prostorsko-časovno nevronsko mrežo z dvojnim tokom in mehanizmom pozornosti, kar izboljšuje prepoznavo košarkarskih akcij s hitrostjo 20 sličic na sekundo.

1 Introduction

Movement analysis in basketball can assist coaches and athletes in tactical development and technical training, etc. By accurately identifying and analyzing the movements of players on the court, the efficiency and quality of game analysis can be further improved to help teams optimize their strategies and enhance their sports performance. However, basketball sports video data often have more complex backgrounds, shorter movement times, and occlusion between players, which poses a higher challenge for basketball sports action recognition [1, 2]. Deep learning (DL) is a considerably more developed technology that is utilized extensively in the field of video image processing. Neural networks, namely convolutional neural networks (CNNs) and recurrent neural networks (RNNs), among others, have proven to be highly effective in processing data from video sequences. Nevertheless, traditional DL models still face limitations in basketball motion analysis, such as RNN's susceptibility to gradient vanishing and explosion, leading to long-term dependency problems. Moreover, although CNNs are outstanding in image recognition, they do not perform well enough in processing time series information [3-5]. Consequently, the field of research has turned its attention to efficient spatio-temporal information fusion to raise the precision and stability of action recognition. Attention mechanism (AM) has emerged as a critical technology in recent years to address the aforementioned issues. By eliminating the interference of unnecessary information, AM helps the model to identify significant features and concentrate on the crucial portions of the input data [6]. Therefore, the study proposes a novel multi-head Sigmoid self-attention (MHSSA) temporal capture module for spatio-temporal feature extraction. Aiming at the temporal correlation of actions in video sequences, it provides a more effective way to recognize actions, and can focus on multiple time points simultaneously, thus capturing the temporal features of actions more comprehensively. In addition, a spatial attention network is designed to capture correlation patterns between spatial targets. This network combines a residual structure and a Sigmoid activation function for generating multi-class attention values. The study is broken down into four sections: The first examines the state of action recognition research at the moment. The second creates the study's proposed two-stream fusion optimization network. The third examines the model's experimental analysis; and the fourth compiles the experimental data.

This year, one of the main study themes is the development and maturation of the field of video action recognition. A kernel attention adaptable graph Transformer network (GTN) was proposed by Liu et al. To capture the higher order spatio-temporal dependencies of joints in skeletal data, they used temporal kernel attention and MHSSA. Additionally, the adaptive approach and dual-flow framework were shown for optimization. On the NTU-RGBD 60 dataset, the results showed that their model was 1.9% more accurate than the baseline 2s-AGCN [7]. Chen et al. suggested a two-stream graph convolution network architecture. This architecture, which combines spectral and vertex-domain graph convolution based on graph Fourier transform, can better detect complex activity and extract spatio-temporal information from skeletal data. In tests on large datasets such as NTU-RGBD and Kinetics-Skeleton, their models performed better [8]. Liu et al. proposed a two-stream cross-modal fusion Transformer action recognition model. The model improved the feature representation and interaction of the three primary colors and depth modalities through self-attention and cross-attention modules. In addition, a bottleneck excitation feedforward block was introduced to enhance the model capability and reduce the computational overhead. Experimental results indicated the effectiveness and generalizability of their action recognition model [9]. Zhong et al. designed a novel multimodal human behavior recognition network by combining a transformer-based skeleton self-attentive subnetwork and a CNN-based deep self-attentive subnetwork. By integrating motion synergy features, this approach achieved a recognition rate of up to 90% or more on NTU RGB+D and UTD-MHAD datasets, outperforming most existing methods [10].

Li et al. proposed a triple-attention module for enhancing the ability of graph convolutional networks to perceive local motion changes in recognizing actions. The tri-attention module was operated on three domains and aggregated global information as a way to capture significant changes in action sequences. Experiments conducted on the NTU RGB-D and Kinetics-Skeleton datasets proved that the tri-attention module worked well to enhance the network model's performance [11]. Lu et al. proposed a video action recognition network that combines spatial features, temporal features, and spatio-temporal dynamics to improve recognition. AlexNet and LSTM were used as the core components, and the data was processed by special spatio-temporal dynamic perception sub-modules. Tests on UCF 101 and HMDB 51 datasets showed that the recognition accuracy of their design model reached 93.53% and 69.36%, respectively [12]. Pau Climent-Pérez et al. used a spatio-temporal attention network, combined with standardized skeletal pose data and RGB data, to significantly improve the recognition of daily activities on the Smarthomes dataset. The method outperformed the existing state-of-the-art by 9.5% and achieved view-invariant action recognition [13]. Gutoski ELHS et al. proposed an open-set human action recognition for action type recognition beyond the training set and developed a novel DL model based on a triple-inflated 3D convolutional network model focusing on feature extraction to distinguish between known and unknown actions. Tests on the UCF-101 dataset showed that TI 3D outperforms other non-metric learning models for action recognition [14]. The classification and statistical results of relevant literature research are obtained through collation, as shown in Table 1.

Table 1: Literature research results

Year	Author	Method	Achievement
2021	Li et al. [11]	Tri-attention enhanced graph convolutional network	Improved skeleton based action recognition
2021	Climent-Perez et al. [13]	Spatio-temporal attention with skeletal and video data	Improved daily activity recognition
2021	Gutoski et al. [14]	Deep metric learning for open-set action recognition	Improved recognition of known and unknown actions
2022	Liu et al. [7]	Graph transformer network with temporal kernel attention	Improved skeleton-based action recognition
2022	Chen et al. [8]	Dual-domain graph convolutional networks	Enhanced action recognition from skeleton data
2022	Liu et al [9]	Dual-stream cross modality fusion transformer	Improved RGB-D action recognition
2023	Zhong et al. [10]	Multimodal cooperative self attention network	High recognition rates for human actions

2023	Lu et al. [12]	Siamese motion-aware spatio-temporal network	Improved video action recognition

Enhancing key features and extracting spatio-temporal features are crucial for video action detection. The aforementioned studies have usually investigated only a part of them and neglected the combination of the two. Therefore, the study improves the spatio-temporal network by MHSSA attention module and residual network (ResNet) to enhance the exercises between frame images. In addition, the two network models are also fused by optimizing the two-stream architecture, aiming to further enhance the recognition analysis of key features.

2 Application of deep learning network incorporating attention mechanism in two-stream architecture for action recognition

For the design of action recognition model for basketball players, the study first proposes a two-stream fusion algorithm and improves it. Subsequently, the temporal domain neural network and spatial domain neural network are optimized by AM and DL optimization, respectively. Lastly, the model's accuracy in recognizing basketball players from film action is improved.

2.1 Network architecture construction based on two-stream fusion algorithm

Basketball is a hand-centered physical confrontational sport, and there are more violations to be judged during the game. For example, a player intentionally kicking the ball with his foot or intercepting the ball with any part of his foot should be judged as a violation, as well as interfering with the rhythm of the shooter by lightly hitting his wrist and other positions at the moment when the opponent shoots the ball is considered a violation. Secondly, basketball players also need to standardize their movements when they train. Violation action recognition and training actions are usually subtler observations, which may be difficult to accurately realize only through the traditional human eye discrimination. Action recognition technology has advanced and is now widely employed in a variety of industries, including sports and education. Basketball players' movement posture can be standardized, evaluation discrimination can be automated, and basketball players' training effects can finally be improved by combining action recognition and basketball [15]. The study uses more mature DL techniques for the design of recognition algorithms. Among them, the two-stream fusion algorithm has less number of parameters and relatively shorter tuning and training time compared to the 3D convolutional model, so the study chooses the two-stream architecture for network construction. Figure 1 shows the flow of the two-stream fusion algorithm.

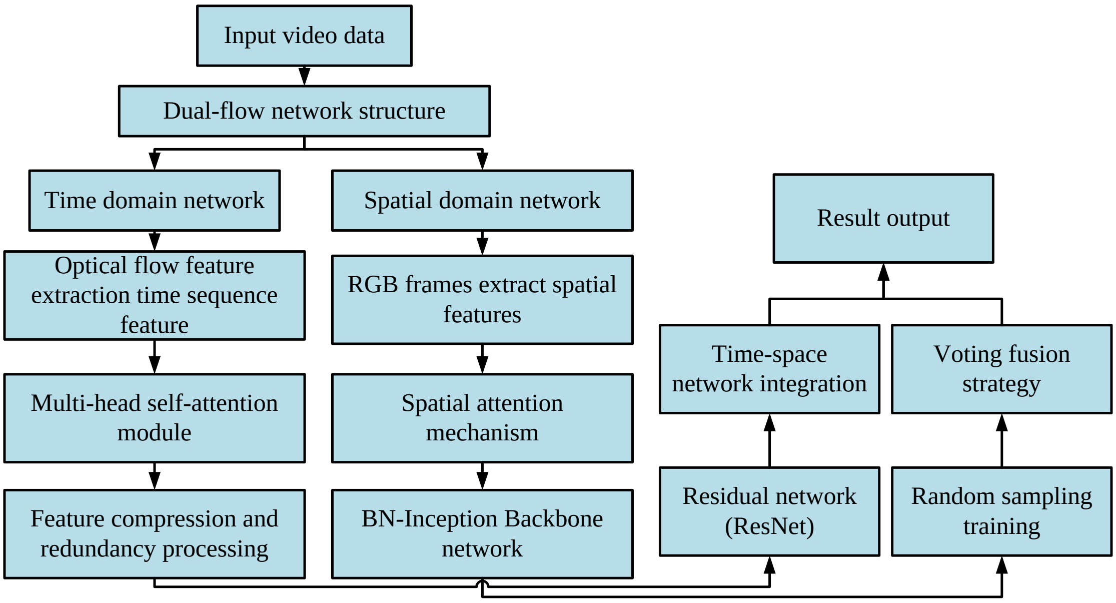

Input video data
Dual-flow network structure
Time domain network
Spatial domain network
Optical flow feature extraction time sequence feature
RGB frames extract spatial features
Multi-head self-attention module
Spatial attention mechanism
Feature compression and redundancy processing
BN-Inception Backbone network
Result output
Time-space network integration
Voting fusion strategy
Residual network (ResNet)
Random sampling training

Figure 1: The flow of the two-stream fusion algorithm

The two-stream algorithm includes two modules in the temporal and spatial domains, the optical stream features are used for the extraction of short-range action timing features, while the RGB frames are used for the extraction of spatial features. When the variation of continuous image frames exceeds 24 frames per second, the presentation effect of video will be reached. This necessitates the use of a digital adjustment of the analog image, as the general algorithm is unable to process the video metadata, which contains a large amount of information. While the basic unit pixels of the simulated image, all contain different gray values and color values, in order to quantify them, the study chooses the RGB quantization method that contains three color channels, red, green and blue, which can be superimposed on the different channels to obtain all the human-perceivable colors, whose descriptions of spatial features are shown in equation (1) [16, 17].

$$img_{RGB} \in \square^{3\times h\times w} \qquad (1)$$

In equation (1), h/w is the height and width of the image, respectively. Compared with the spatial features of video images, they are more strongly connected in temporal features. If the extraction of temporal features is performed only by RGB data, it is very easy to have problems such as excessive computational burden and overfitting. Therefore, it is investigated to use the optical flow features to compute the variation of pixel point intensities with respect to time frames and to obtain the moving orientation and velocity magnitude of the recognized object. Equation (2) illustrated the optical $I\left(x,y,t_1\right)$ of a pixel in a certain frame.

$$I\left(x,y,t_1\right)=I\left(x+\Delta x,y+\Delta y,t_1\right) \qquad (2)$$

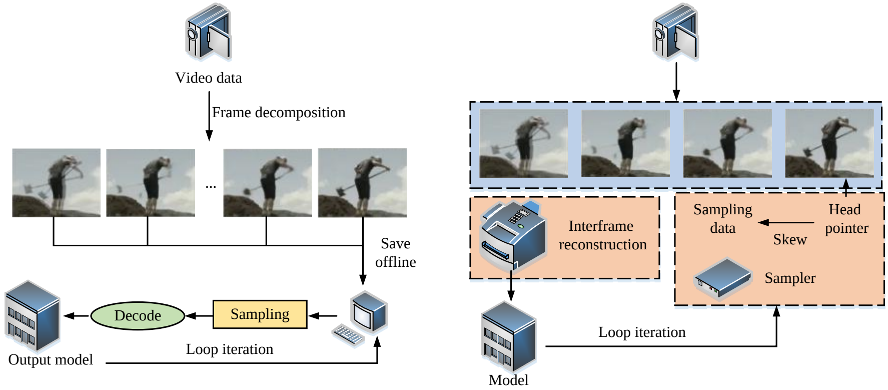
In equation (2), $\left(x,y\right)$ denotes the pixel position at the t_1 moment and $\left(\Delta x,\Delta y\right)$ denotes the pixel movement at the next moment. When the amount of movement is small, its optical flow characteristics can be obtained by Taylor series as shown in equation (3) [18, 19].

$$I_xV_x+I_yV_y+I_t=0 \qquad (3)$$

In equation (3), V_x/V_y denotes the optical flow velocity in different orientations, and $I_x/I_y/I_t$ denotes the bias of the optical brightness function in the $x/y/t$ direction. However, too large a gap in pixel movement in an image frame tends to lead to training oscillations, so the study first truncates the extreme optical flow and unifies the optical flow values by Min-Max normalization as shown in equation (4).

$$o_{new}=255\times\frac{o-o_{\min}}{o_{\max}-o_{\min}} \qquad (4)$$

In equation (4), $o_{\max}/o_{\min}$ is the very large and very small values of the optical flow, and o/o_{new} is the optical flow value before and after normalization, respectively. The traditional two-stream architecture is to directly split and save and sample the images of each frame, which produces large redundant data, aggravates the storage space requirement, and ultimately leads to a reduction in the real-time performance of the model. Therefore, the study firstly compresses and saves the redundant information, and the data includes the key frame I with simple compression, the prediction frame P containing the temporal relationship, and the bidirectional prediction frame B with bidirectional temporal search [20-23]. The traditional and improved architectures are shown in Figure 2.



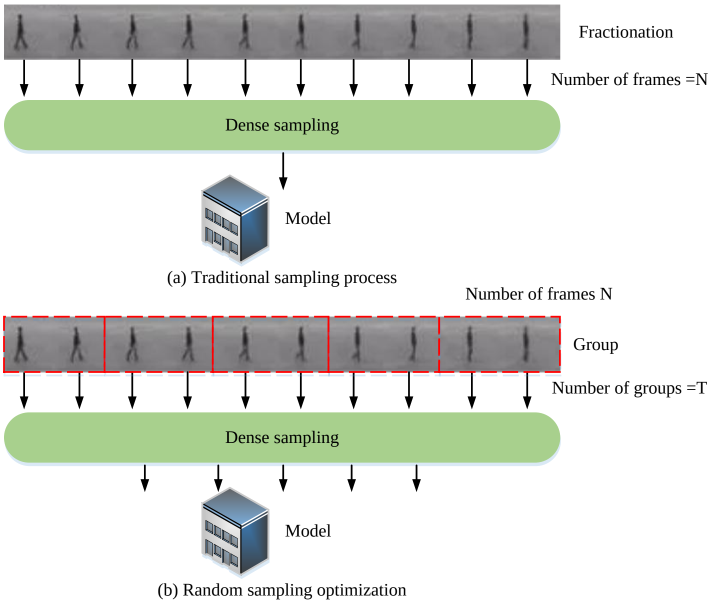
(a) Traditional Two-Stream architecture (b) Optimize the Two-Stream architecture

Figure 2: Two-stream architecture optimization

Random sampling training reduces the computational burden on the model and enhances the efficiency of modeling temporal information. The study uses this grouped form of sparse sampling for spatial-domain network training, which first requires an equal segmentation of the image frames for independent sampling of each group before inputting the splicing results into the value-space network. The sampling of the time-domain network is performed simultaneously in both directions, so the network input values are continuous stacked optical flow groups. The inputs to the two neural spaces are shown in equation (5).

$$\begin{cases} Input_{Spatial} = Concat\left(\underbrace{\{img_{v_1}, img_{v_2}, ..., img_{v_T}\}}_{T}\right) \\ Input_{Temporal} = concat\left(\{O^u_{t_0}, O^v_{t_0}, ..., O^u_{t_0+k-1}, O^v_{t_0+k-1}\}\right) \end{cases} \quad (5)$$

In equation (5), T denotes the number of video cut groups and N denotes the number of video image frames. The video is denoted as $V = \{img_1, img_2, ..., img_N\}$ and t_0 / k denotes the sampling starting point and the number of stacking groups. $O^u_{t_0} / O^v_{t_0}$ denotes the optical flow characteristics of the starting point t_0 image in the horizontal and vertical directions, respectively. When the traditional frame directly segments the image, the overfitting problem due to the similarity of each frame occurs, while the above sampling can effectively solve this phenomenon, as shown in Figure 3.



(a) Traditional sampling process

(b) Random sampling optimization

Figure 3: Sampling method optimization

The neural network of two-stream architecture consists of a backbone network for global action feature extraction, and a recognition layer for judgment. The common batch normalization-inception (BN-Inception) backbone network is selected for the study, which serves as an updated version of the inception network with a regularization layer that better achieves data stabilization and improves the original convolutional kernel, as shown in Figure 4.

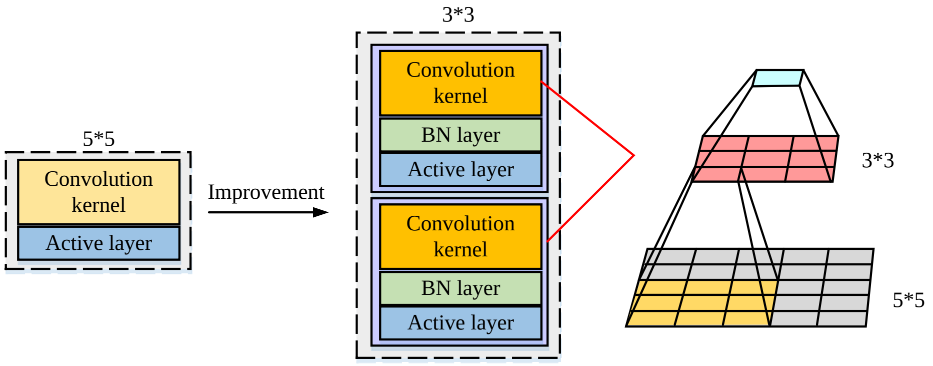


Figure 4: Backbone network structure of BN-Inception

The traditional Inception network with 5*5 convolutional kernels is replaced with 3*3 convolutional kernels. The increase in model depth also provides a better environment for detailed feature extraction. Moreover, the sensory field of the model remains stable in the stacked convolutional layers. The input value attributes of the time-space domain neural network are different. For example, if the number of input channels of the time-domain network is 2k, the initial convolutional kernel substitution should be expressed as 2k→64 and the parameters should be initialized to the original parameter averages as shown in equation (6) [24].

$$ConvPara_i^{Temporal} = \frac{1}{3}\sum_{j=1}^{3} ConvPara_i^{Spatial} \qquad (6)$$

In equation (6), $ConvPara_i^{Temporal} / ConvPara_i^{Spatial}$ is the initial convolutional parameter of the spatio-temporal domain neural network, respectively. Subsequent BN layers are able to strengthen the generalization performance of the network, and Partial BN module is introduced to lock the regularization layers other than the first BN layer, which are normalized by the pre-training parameters. Finally, the fusion of the time-space domain network is carried out by utilizing the voting fusion strategy, as shown in equation (7).

$$P = \sum_{i=1}^{2} Weight_i * P_i \qquad (7)$$

In equation (7), P_i denotes the valuation of video image frame i by the space-time-domain network. $Weight_i$ denotes the weight of video image frame i.

2.2 Time-space domain action recognition algorithm based on attention mechanism and deep learning network

The optical flow features in the infrastructure only enable the extraction of temporal relationships over short periods of time, whereas the movements of basketball players tend to be of longer duration. Moreover, the weights of the images in each frame are not consistent, so it is also not possible to sample the mean or extreme value for the extraction of time-domain features. RNNs have been widely used in the field of action and timing prediction as models with high dynamic temporal description [25, 26]. However, the excessive similarity between image frames can limit the optimization process of the network, and the computational dependency between its timing modules can slow down the convergence of the model. Therefore, the study selects AM in DL for the design of the timing capture module. The performance of the model for the detailed features will be weakened by the traditional encoder-decoder's interpretation of the data as independent features. With the introduction of AM, the model is able to implement different attention to data with different weights, as shown in Figure 5.

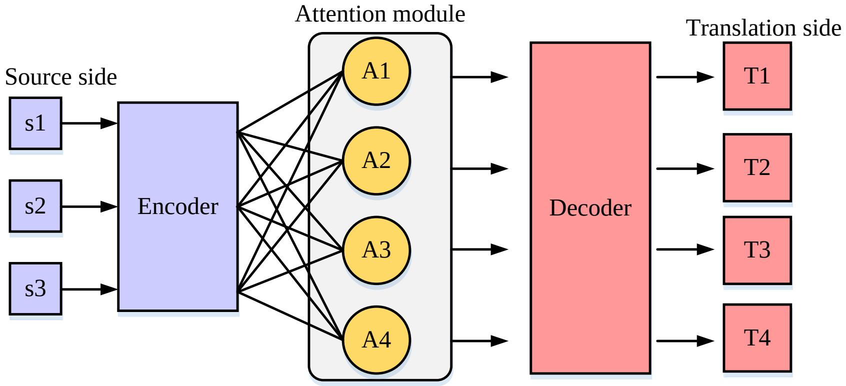


Figure 5: Introduction structure of attention mechanism

As illustrated in Figure 5, the extraction of optical flow features in the infrastructure is limited to a brief period. To enhance the time-domain neural network, an AM is incorporated. Through linear combination of diverse weights, the AM translates the dimensionality of source data to the translation end, optimizes the accuracy of feature extraction through continuous training iterations, and enhances the recognition effect of the model in time series data. The source data is mapped on the translation side by a linear combination of different weights and dimensionality reduction, as shown in equation (8).

$$T\arg et_j = Embedding\left(\sum_i a_{ji} * Source_i\right) \qquad (8)$$

In equation (8), $Source_i / T\arg et_j$ denotes the source code end and the translation end respectively, and $Embedding$ is the dimension reduction layer. a_{ji} is the corresponding weights of the translation end and the source code end, which are obtained after continuous training iterations. AM is essentially a query module for the data, i.e., the data values of the translation end and the source code end are similarity computed in order to obtain the corresponding weights, which are then transformed into the probabilistic model a_i in Softmax as shown in equation (9) [27].

$$a_i = Soft\max(e_i) = \frac{\exp^{e_i}}{\sum_j^n \exp^{e_j}} \qquad (9)$$

In equation (9), e_i is the similarity between the key vector at the source side and the query vector at the translation side, and $\exp^{e_i} / \exp^{e_j}$ is the corresponding primary key value and query vector value, respectively. Finally, the summation is performed according to the weights. The aforementioned AM is unfolded by the Softmax activation function, which has a detrimental effect on the competition among features. It exerts a significant limiting influence on data other than extremely large values and is unable to adapt to the recognition of video actions with a tight temporal connection [28]. Therefore, the study introduces the MHSSA module as shown in Figure 6.

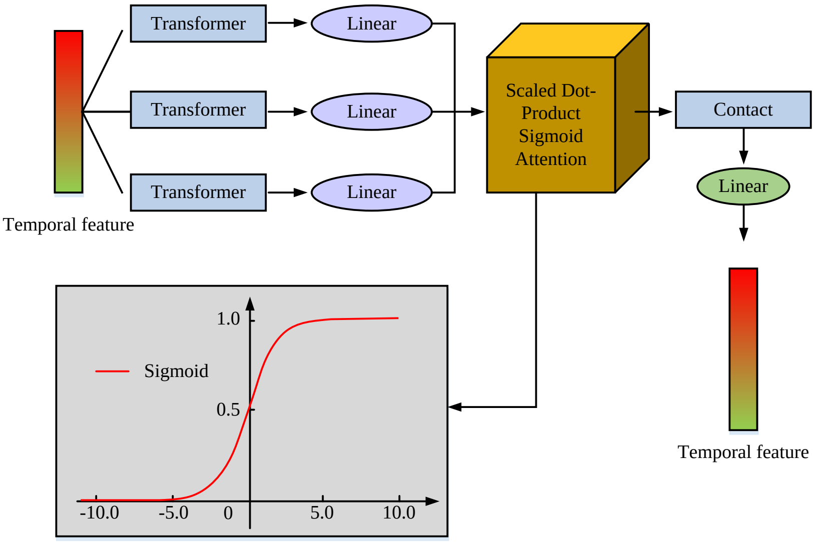


Figure 6: Multi-head Sigmoid self-attention module design

In Figure 6, the Softmax activation function in the traditional AM is replaced with the Sigmoid function. Unlike the original function for direct comparison of temporal features, the Sigmoid function is able to realize the activation of individual neurons through multiple gating functions. In the Sigmoid function curve, the input mapping interval is [0,1], and the overall change is relatively smooth. When the function value is infinitely close to 0, the data obtained is substantially reduced, which may be due to the screening of the background frames in the image to achieve effective control of the data flow. Secondly, the self-AM in the MHSSA module also improves the feature enhancement effect, extracts the temporal relationship between the upper and lower frames, and completes the construction of the action time-domain network. Moreover, the multi-group module assists the self-attention module, which perfects the extraction of temporal feature interrelationships, and then finally activates the neurons through the Sigmoid function. As shown in equation (10) [29].

$$Attention\left(K_i, Q_i, V_i\right) = Sigmoid\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \times V_i \quad (10)$$

In equation (10), $K_i / Q_i / V_i$ denotes the key vector, query vector, and value vector, d_k denotes the key feature dimension, and K_i^T denotes the key matrix, respectively. The MHSSA module employs the Sigmoid function in lieu of the Softmax function for the calculation of attention weights, thereby enhancing the capture of timing features. Equation (11) depicts the specific formula.

$$MHSSA\left(Q, K, V\right) = Sigmoid\left(\frac{QK^T}{d_k}\right) \times V \quad (11)$$

The Sigmoid function is able to smoothly map input values to between 0 and 1, thus effectively regulating the data flow and avoiding excessive restrictions on non-critical features. The multi-head AM divides the input data into multiple sub-spaces, applies the AM to each sub-space independently, and finally concatenates the output of each sub-space, as shown in equation (12).

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W_o \, (12)$$

In equation (12), $head_i$ represents the attention head, and the calculation process is shown in equation (13). Finally, the output of multiple attention heads is combined through concatenation operation, and then multiplied by an output weight matrix W_o to get the final result. After the output of the multi-head AM is spliced and fused, the final eigenvalue is obtained as shown in equation (13).

$$Output = Concat(head_1, head_2, \ldots, head_h)W_o \, (13)$$

In this way, the results of multiple attention heads can be synthesized to improve the feature extraction ability of the model. After obtaining each temporal feature, it is necessary to utilize splicing fusion to obtain the final feature value $Feat_{attn}$, as shown in equation (14).

$$Feat_{attn} = W^{O} \times Concat\left(Attn_1, Attn_2, ..., Attn_{n_{head}}\right) \quad (14)$$

In equation (14), W^{O} denotes the projection matrix and n_{Head} denotes the total number of Heads. The introduction of AM can enhance the extraction of temporal features by the action recognition system. The study proceeds with the construction of a spatial domain network to further enhance the accuracy of basketball motion recognition (BMR). The spatial network aims to unify the recognition of connected targets. For example, during horseback riding, the necessary elements present are the person and the horse, while during basketball movement, the necessary elements are the person and the basketball. In the whole image frame, different recognition targets correspond to different weights and naturally have different contribution values in action recognition. The traditional DL framework uses global average pooling to directly compress the spatial domain features, ignoring the relationship between the spatial domain image frames, so the study optimizes the pooling layer to enhance the extraction of spatial domain features by the model [30]. Global feature extraction takes place in the pooling layer. Two popular pooling techniques are mean pooling and maximum pooling. The former demands that the location of very big values be stored at runtime and used to backpropagate the gradient. The latter requires only the solution of the mean value of the gradient before backpropagation can be passed again. The global pooling layer has the same dimensions as the final layer of the feature layer. Of course, the fusion of the two pooling methods can also enhance the flexibility of the spatial feature module and strengthen the target recognition accuracy of dynamic video, which is especially suitable for multi-target recognition tasks, and the specific module design is shown in Figure 7.

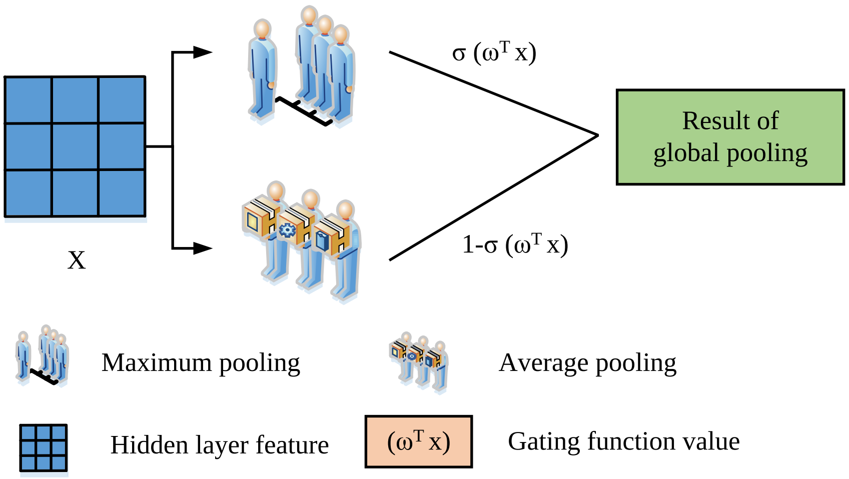


Figure 7 Fusion of average pooling and maximum pooling

Figure 7 illustrates the application of global average pooling and maximum pooling fusion in spatial feature extraction. Global pooling compresses spatial features, yet traditional methods fail to consider the relationship between image frames, resulting in sub-optimal feature extraction. The combination of maximum pooling and mean pooling enhances the flexibility of spatial feature modules, particularly in multi-object recognition tasks. This method significantly improves the recognition accuracy of the model. Action recognition researchers have extensively researched optimization techniques for the pooling layer, such as low-rank approximation based on second-order pooling, etc., which effectively increases the computational efficiency of the model while simultaneously enhancing the capture effect of spatial-domain features through the use of gesture information, as demonstrated in equation (15).

$$f\left(X\right) = a^{T}\left(X^{T}\left(Xb\right)\right) \quad (15)$$

In equation (15), $f\left(X\right)$ denotes the pooling function, $X \in \square^{n\times c}$, $a,b \in \square^{f\times 1}$. Others are augmented by DL network-assisted spatial-domain attention and applied to the feature extraction of the latter frame of the image. The video data is cut off frame by frame over the time series, and the connections between different targets in each image frame are very strong. For example, in the action recognition video of a basketball player, there are often more interactions between the basketball player and the basketball, which means that the two targets are more related. However, image frames not only contain target data, but also usually contain a lot of unnecessary background data. Therefore, the common global pooling approach cannot achieve intelligent partition feature extraction and cannot complete the segmentation of different target data. In more research results have shown that the introduction of neural network module can have an enhanced positive effect on the spatial feature AM. As a result, as seen in Figure 8, the study applies it to the spatial domain module.

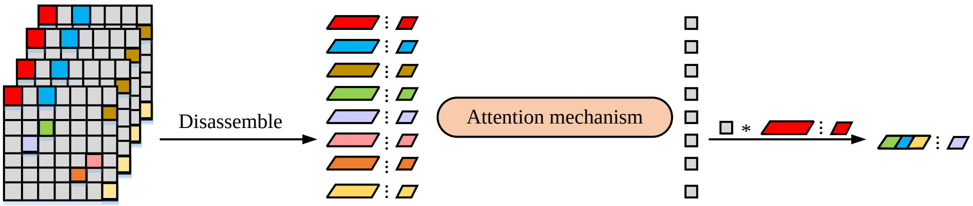


Figure 8: Space and network integrated into attention mechanisms

The spatial information of each frame has a correlation between them, composing their corresponding independent score data, and obtaining the more important spatial region data. The backbone network of the model is BN-Inception, which directly performs correlation modeling when the feature dimension reaches 49*1024. Its need for a large number of pairs will limit the convergence iteration speed of the model and reduce its operational efficiency. Therefore, the study introduces a multi-layer perceptron strategy for improvement. The common activation function itself is competitive and therefore highly susceptible to problems such as inhibition. To solve this problem, the study introduces ResNet to optimize the Sigmoid function. As the number of neural network layers increases, the gradient disappearance or gradient explosion phenomenon is likely to occur. This can be mitigated by introducing the phenomenon into the regularization layer for optimization. However, this approach is not applicable in cases where there are too many layers. At this time, the training set loss will gradually reduce, and tend to saturation, if the number of layers continue to increase, the loss value will increase. As a result, in the event of a network degradation, the shallow network learns more quickly than the deep network, which depends on the residual link to transfer low-level characteristics to higher levels. In deep neural networks, residual connections are frequently used to address the issue of gradient disappearance. They can also enhance the network's training efficiency and model performance, as demonstrated by equation (16).

$$h(x) = f(x) + x \qquad (16)$$

In equation (16), $h(x)$ denotes the constructive formula and $f(x) = x$ denotes the constant mapping relation. The model gradient is smoother in the directly connected line.

3 Time-space domain neural network based on two-stream architecture

Experiments are conducted on two datasets: HMDB51 and Charades-STA. HMDB51 contains 6,766 videos, while Charades-STA contains 6,672 videos. HMDB51 is divided into 51 action categories. Pre-processing steps include data cleaning, noise reduction, and invalid data removal to ensure data quality. Data diversity is enhanced through random cropping, rotation, and scaling. The optical flow algorithm is used to extract the motion timing feature, capturing the direction and speed of pixels over time, and extracting the RGB frame of the video to obtain the spatial feature. Min-max normalization is used to compress and store key frames, prediction frames, and bidirectional prediction frames to reduce redundancy and improve real-time performance. The initial learning rate is 0.001, the batch size is 128, and 80 iterations are performed. Simulation studies are carried out to assess the study design model for BMR recognition in order to confirm its efficacy. Firstly, the key modules in the time-domain neural network and spatial neural network, as well as the overall performance are analyzed respectively. Finally, the overall model after fusion is subjected to performance comparison and analysis experiments in different datasets.

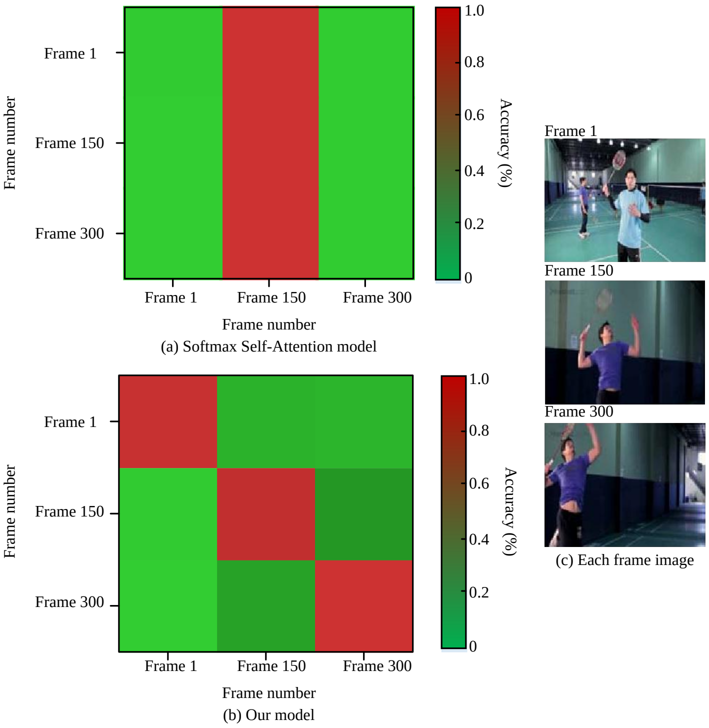
3.1 Performance analysis and validation of time-domain neural network and spatial neural network

The study begins with a performance validation analysis of the time-domain neural network as well as the spatial domain neural network respectively. The experimental environment and parameter settings are shown in Table 2.

Table 2: Experimental environment and parameter Settings

Name	Settings
Operating system	Ubuntu 16.04.2 LTS
Internal memory	64GB
CPU	InterR CoreTM i7-5930K CPU
Graphics card	GeForce GTX TITAN Xp
Deep learning framework	PyTorch
Data sampling pre-processing	TSN open source
Data set	HMDB51 Charades-STA
Batch size	128
Equal sampling values	24
Initial learning rate	0.001

The HMDB51 dataset contains a total of 6766 video data, with 51 action categories, including several confusing action categories such as handstand as well as back handspring, and the ratio of training to validation is 7:3. The study first experimentally analyzes the attentional module in the time-domain attentional network, and compares it with the common Softmax self-AM. Moreover, among the video data, the first, 150th, and 300th frames are chosen for examination. Figure 9 displays the outcomes of the experiment.



(a) Softmax Self-Attention model

(b) Our model

(c) Each frame image

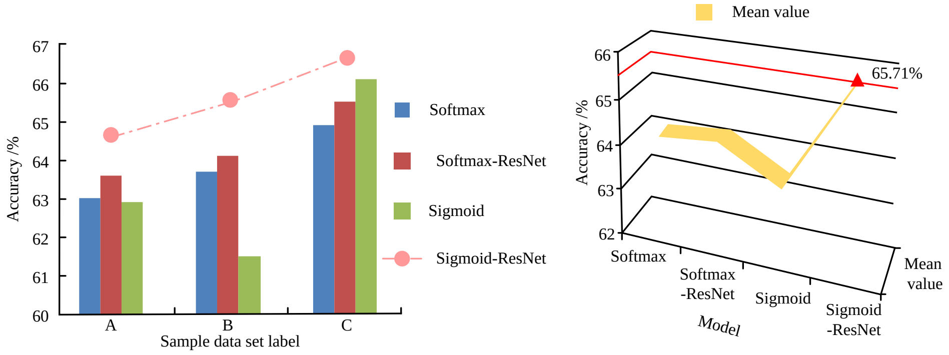
Figure 9: Comparison of performance of attention module in time-domain network

Figure 9(c) shows the images of each frame used in the experiment, and the sampled data is the video of playing badminton action. In Figure 9(a), the Softmax self-AM module performs poorly, and the recognition performance is better only in the 150th frame image, with an accuracy of 99.9%, while for the rest of the frames, the recognition error is very large, and the recognition accuracy is even below 1%. This is due to the overly competitive mechanism of the Softmax function and the fact that its attention module does not take into account multi-frame data variations, but rather focuses on independent frame images, limiting the contribution of the remaining frames. In the study of the proposed MHSSA module, its temporal recognition performance is significantly improved, with recognition accuracies all above 85%, averaged over a more focused combination of multi-temporal frame images. The study compares the time-domain MHSSA model with different number of multi-modules, with RNN, and its common variant forms long short term memory network (LSTM), and gated recurrent unit (GRU), and the experimental results are shown in Figure 3.

Table 3: Time sequence feature extraction accuracy for each model action recognition

Model	Sample set sequence number			Mean value
	A	B	C	
RNN	52.81%	53.74%	55.23%	53.93%
LSTM	53.47%	53.99%	56.05%	54.50%
GRU	53.92%	54.68%	56.88%	55.16%
1 Head	63.45%	64.23%	65.17%	63.35%
2 Head	64.57%	63.72%	66.98%	65.09%
4 Head	63.01%	64.60%	66.95%	64.85%
6 Head	63.36%	63.24%	65.71%	63.92%

In Table 3, neural networks such as RNN are significantly worse than the time-domain neural networks proposed in the study, and their recognition accuracies are all under 60%. The LSTM model and GRU model perform slightly better than the RNN model due to the additional implicit states added by the LSTM model, which can better solve problems such as gradient explosion compared to the RNN network. The GRU network, although similar in structure to the LSTM network, has a simpler structure and utilizes its iterative updates. However, in motion video, the high similarity of each image frame makes it difficult for all three models mentioned above to obtain suitable learning patterns, and thus the overall recognition results are poor. The action recognition of the time-domain MHSSA model proposed in the study is significantly improved, with the highest model recognition accuracy in sample dataset C for all module numbers, reaching 66.20% on average. When the number of modules is 2, the overall model has the best recognition performance, which is higher than the rest of the MHSSA models by 1.91% on average, and higher than the three neural network models by 10.56% on average. In conclusion, it can be concluded that the proposed MHSSA-based temporal feature extraction module of the study has significantly improved the action recognition network. The study conducts experiments to analyze the performance of spatial neural networks with the same principle. The residual module in the model is first validated and Softmax and Sigmoid functions are chosen for comparison respectively. Figure 10 displays the outcomes of the experiment.



(a) Performance of each model in different sample sets

(b) Average recognition accuracy of each model

Figure 10: Spatial neural network module performance analysis

In Figure 10(a), the addition of the ResNet module to both activation functions resulted in an increase in the model's recognition performance. In particular, the network model recognition accuracy based on the Softmax function increased by 0.54% on average, and the network model recognition accuracy based on the Sigmoid function increased by 2.10% on average. As shown in the model comparison between the Softmax and Sigmoid functions, the Sigmoid function used in the study is almost always due to the Softmax function before and after optimization. Moreover, only in sample dataset B, it is lower than the Softmax model and

Softmax-ResNet model by 2.24% and 2.63%, respectively. The Sigmoid-ResNet model, on the other hand, always has the highest recognition accuracy and the performance improves as the dataset changes. In dataset A, the accuracy is improved by 1.71% compared to the Sigmoid function model, and in sample dataset B, it is improved by 4.0%. Figure 10(b) shows the mean value of the accuracy of each model, excluding the model used in the study, the mean value of the accuracy of the remaining models is below 65%. While the Sigmoid function model has the lowest mean value, the optimized Sigmoid-ResNet model reaches the maximum value of 65.71%, which is higher than the rest of the models by 1.9% on average. Finally, the study fused the time-domain neural network and spatial neural network through two-stream architecture, and the iteration results before and after model optimization are shown in Figure 11.

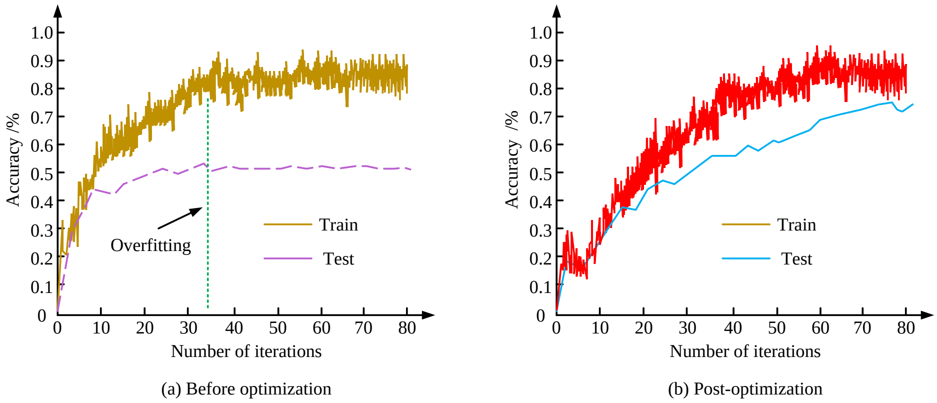


Figure 11: Comparison of iteration results before and after model optimization

In Figure 11(a), the pre-optimization action recognition model suffers from overfitting around the 35th iteration and fails to achieve a more stable DL. In contrast, in Figure 11(b), the design model optimized by fusion of the two-stream architecture solves the overfitting phenomenon present in the initial network. In the 80 iterations of the result intercept, the model is always learning iterations and does not converge prematurely. In summary, it can be seen that the research design model enables more stable and effective learning optimization.

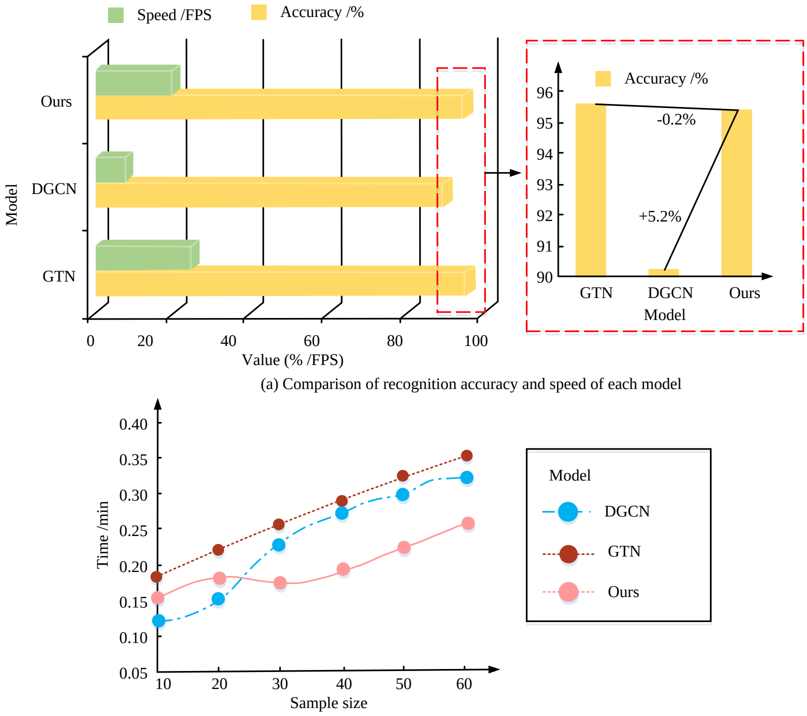
3.2 Performance analysis of spatio-temporal neural network model action recognition under two-stream architecture fusion

The study further conducted performance analysis experiments on the fused overall model and selected the Charades-STA dataset and ActivityNet-Captions dataset as the experimental samples. There are a total of 6672 videos in the Charades-STA dataset, and the duration of each video data is about 30s. The ActivityNet-Captions dataset, as a larger dataset, covers a more diverse range, and the total duration of each video is around two minutes. The study divides the training set, validation set and test set of the experiment in a 2:1:1 manner. After that, the evaluation metric "R@n, IoU=m" is chosen, which aims to determine the proportion of samples that the corresponding model recognizes the first n image frames with no less than one successful recognition. Where, not less than 1 means that the intersection over union (IoU) between the judgment value and the true value is not less than m. Finally, the study introduces the optimized GTN proposed by Liu et al. and dual-domain graph convolutional networks (DGCN) proposed by S Chen et al. for controlled experiments. Table 4 displays the findings of the experiment.

Table 4: Performance comparison of different models in different data sets

Index	Data sets					
	Charades-STA			ActivityNet-Captions		
	Model			Model		
	GTN	DGCN	Ours	GTN	DGCN	Ours
R@1, IoU=0.5	45.40	42.61%	53.07%	42.48%	43.97%	45.46%
R@1, IoU=0.7	26.57	25.80%	31.79%	22.26%	23.81%	24.39%
R@5, IoU=0.5	88.03	79.42%	89.11%	71.84%	67.93%	78.02%
R@5, IoU=0.7	55.39	54.88%	60.24%	45.97%	50.03%	50.79%

In Table 4, in Charades-STA dataset, the best performance of the research design model, excluding the R@1, IoU=0.7 case where both mean values are lower, reaches 67.47%, which is on average 7.86% higher compared to the remaining two models. Moreover, the same result is observed in the ActivityNet-Captions dataset. The mean value of the research design model reached 58.09% when the extremes are discarded, which is elevated by 6.92% on average compared to the remaining two models. Moreover, when the IoU is 0.5, the performance of the models are due to the case of IoU of 0.7. The study is further analyzed experimentally for BMR recognition and the results are shown in Figure 12.



(a) Comparison of recognition accuracy and speed of each model

(b) Effect of sample size on model recognition time

Figure 12: Comparison of basketball action recognition performance of each model

In Figure 12(a), the DGCN model has a relatively low accuracy despite its fast recognition speed. Moreover, the recognition speed of the research design model is medium, reaching 20 FPS, which is 25% lower than that of the GTN model. Although its recognition accuracy is slightly lower than the GTN model by 0.2%, overall, the research design model has the best overall performance. In Figure 12(b), a shooting batter's action is selected and experimented separately with different number of samples. The experimental results show that the recognition time of each model increases with the sample size. The GTN model has the longest recognition time with an average of 0.269 minutes, while the DGCN model initially has the shortest recognition time, but it gradually increases with the sample size and surpasses the research design model after 20 samples. The average recognition time of the research design model is, on average, 47.27% lower than the rest of the models. In conclusion, it can be concluded that the research-designed optimized spatio-temporal domain neural network based on two-stream architecture is able to achieve BMR recognition better. To further verify the performance of the research method, the research method is compared with the existing motion recognition techniques. Descriptive statistical analysis calculates mean ± standard deviation. Independent sample t test is used to evaluate the significance, and P<0.05 is considered statistically significant. The performance comparison results of various action recognition methods are shown in Table 5.

Table 5: Performance comparison results of various action recognition methods

Method	Recall rate (%)	F1 score	Model complexity (%)	Accuracy rate (%)
Research method	98.11±0.35	95.50±0.96	54.35±0.26	98.24±0.34
Literature [7]	91.29±0.41	89.74±0.74	58.39±0.77	87.27±0.36
Literature [8]	86.34±0.39	87.69±0.58	60.12±0.41	89.64±0.73
Literature [10]	87.28±0.47	92.67±0.14	67.63±0.34	87.65±0.65
Literature [13]	89.24±0.36	91.73±0.21	70.24±0.37	84.76±0.76
P	<0.05	<0.05	<0.05	<0.05

Table 5 presents the results of the motion recognition method under study, demonstrating superior performance in terms of recall rate, F1 score, and accuracy, exceeding 93%. In comparison to existing pose recognition methods, the research method exhibits a significantly lower model complexity, with a value of only 54.35%. The P values of all comparison results are less than 0.05, indicating that the experimental results are statistically significant.

4 Discussion

Compared with the traditional Softmax mechanism and the RNN/LSTM/GRU based model, the MHSSA module achieved a higher accuracy in time series feature extraction, with an average recognition accuracy of more than 85%. On HMDB51 dataset, the basketball action recognition accuracy reached 95.4%, and the comprehensive performance on Charads-STA and ActivityNet-Captions dataset was 7.86% and 6.92% higher than that of GTN and DGCN models, respectively. In addition, the recognition speed of this method was also significantly improved, and the average recognition time was reduced by 47.27%. Compared with Liu et al. 's GTN [7] and Chen et al.' s two-flow graph convolutional network [8], the innovation in this study was the introduction of multi-head self-attention and spatial AMs to enhance the ability to capture spatio-temporal features. By optimizing the two-stream architecture, the features of the time-domain and the space domain were effectively integrated to improve the overall performance and efficiency. Conventional models are inadequate for handling lengthy time series and intricate spatial characteristics. This is where the AM and architectural optimization come into play. The model combines advanced DL techniques and AMs to achieve more efficient feature extraction and fusion through multi-layer perceptrons and ResNet optimization.

5 Conclusion

In order to enhance the practice effect of basketball players, the study proposes an action recognition model based on DL network. Aiming at the temporal correlation of actions in video sequences, it provides a more effective way of action recognition, which is capable of focusing on multiple time points at the same time. In addition, a spatial attention network is introduced for capturing association patterns between spatial targets. Experiments were conducted to analyze the temporal neural network and spatial neural network respectively. In the former experiment, the results indicated that Softmax self-AM performs poorly in action recognition due to single centralization, and the recognition accuracy in other frames was even less than 1%. In contrast, the proposed MHSSA module achieved a recognition accuracy of over 85%, which is significantly better than the Softmax mechanism. Furthermore, on average, the MHSSA model achieved a recognition accuracy 10.56% higher than that of the neural network model when compared to traditional RNN/LSTM/GRU networks. The experimental results of the spatial neural network indicated that the model with the inclusion of the residual module improved the accuracy by 1.71% compared to the Sigmoid letter model in the dataset A. The model with the inclusion of the residual module improved the accuracy by 1.71% compared to the Sigmoid letter model. The fusion of the spatio-temporal network through the two-stream architecture effectively solved the overfitting problem in the original network. In the Charades-STA dataset, except for the case of R@1, IoU=0.7, the average performance of the model reached 67.47%, which is 7.86% higher than that of the GTN and DGCN models. In the ActivityNet-Captions dataset, the average performance of the model reached 58.09%, which is 6.92% higher than other models. For BMR recognition, the research-designed model achieved moderate recognition speed and only slightly lower recognition accuracy than the GTN model by 0.2%, with the best overall performance. In the experiments with different sample sizes, the average recognition time of the research-designed model was 47.27% lower than the average. In summary, it can be concluded that the research design model is able to perform BMR recognition more accurately and quickly. However, the above study is based on the two-bit network, and further extraction should be carried out subsequently for the three-dimensional spatio-temporal features. The current research is based on a two-dimensional action recognition network, which can be extended to three-dimensional spatio-temporal feature extraction in the future to improve the recognition effect of complex actions. This model can be applied to basketball training and game analysis to accurately identify and analyze players' movements. It can also be combined with AI technology to help coaches develop personalized training plans.

References

[1] D. Lee, D. Wang, Y. Yang, L. Deng, and G. Li, "QTTNet: Quantized tensor train neural networks for 3D object and video recognition," Neural Networks, vol. 141, no. 5, pp. 420-432, 2021. https://doi.org/10.1016/j.neunet.2021.05.034

[2] M. Smith, and R. Toumi, "Using video recognition to identify tropical cyclone positions," Geophysical Research Letters, vol. 48, no. 7, pp. 1-9, 2021. https://doi.org/10.1029/2020GL091912

[3] A. M. Atto, A. Benoit, and P. Lambert, "Timed-image based deep learning for action recognition in video sequences," Pattern Recognition, vol. 104, no. 1, pp. 107353-107366, 2020. https://doi.org/10.1016/j.patcog.2020.107353

[4] N. Hajarolasvadi, and H. Demirel, "Deep facial emotion recognition in video using eigenframes," IET Image Processing, vol. 14, no. 14, pp. 3536-3546, 2020. https://doi.org/10.1049/iet-ipr.2019.1566

[5] Y. Wu, Y. Fu, and S. Wang, "Deep instance segmentation and 6D object pose estimation in cluttered scenes for robotic autonomous grasping," Industrial Robot, vol. 47, no. 4, pp. 259-273, 2020. https://doi.org/10.1108/IR-12-2019-0259

[6] A. Mirza, and I. Siddiqi, "Recognition of cursive video text using a deep learning framework," IET Image Processing, vol. 14, no. 14, pp. 3444-3455, 2020. https://doi.org/10.1049/iet-ipr.2019.1070

[7] Y. Liu, H. Zhang, D. Xu, and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," Knowledge-Based Systems, vol. 240, no. 3, pp. 108146-108152, 2022. https://doi.org/10.1016/j.knosys.2022.108146

[8] S. Chen, K. Xu, Z. Mi, X. Jiang, and T. Sun, "Dual-domain graph convolutional networks for skeleton-based action recognition," Machine Learning, vol. 111, no. 7, pp. 2381-2406, 2022. https://doi.org/10.1007/s10994-022-06141-8

[9] Z. Liu, J. Cheng, L. Liu, Z. Ren, Q. Zhang, and C. Song, "Dual-stream cross-modality fusion transformer for RGB-D action recognition," Knowledge-based Systems, vol. 255, no. 11, pp. 109741-109752, 2022. https://doi.org/10.1016/j.knosys.2022.109741

[10] Z. Zhong, Z. Hou, J. Liang, E. Lin, and H. Shi, "Multimodal cooperative self-attention network for action recognition," IET Image Processing, vol. 17, no. 6, pp. 1775-1783, 2023. https://doi.org/10.1049/ipr2.12754

[11] X. Li, W. Zhai, and Y. Cao, "A tri-attention enhanced graph convolutional network for skeleton-based action recognition," IET Computer Vision, vol. 15, no. 2, pp. 110-121, 2021. https://doi.org/10.1049/cvi2.12017

[12] X. Lu, W. Quan, R. Marek, H. Zhao, and J. X. Chen, "SiamMAST: Siamese motion-aware spatio-temporal network for video action recognition," The Visual Computer, vol. 1, no. 1, pp. 1-19, 2023. https://doi.org/10.1007/s00371-023-03018-2

[13] P. Climent-Pérez, and F. Florez-Revuelta, "Improved action recognition with separable spatio-temporal attention using alternative skeletal and video pre-processing," Sensors, vol. 21, no. 3, pp. 1005-1005, 2021. https://doi.org/10.3390/s21031005

[14] M. Gutoski, A. E. Lazzaretti, and H. S. Lopes, "Deep metric learning for open-set human action recognition in videos," Neural Computing & Applications, vol. 33, no. 4, pp. 1207-1220, 2021. https://doi.org/10.1007/s00521-020-05009-z

[15] S. Nag, P. Shivakumara, U. Pal, T. Lu, and M. Blumenstein, "A new unified method for detecting text from marathon runners and sports players in video," Pattern Recognition, vol. 107, no. 1, pp. 107476-107476, 2020. https://doi.org/10.1016/j.patcog.2020.107476

[16] B. Tao, and P. Toumi, "Attention-based LSTM-FCN for earthquake detection and location," Geophysical Journal International, vol. 228, no. 3, pp. 1568-1576, 2021. https://doi.org/10.1093/gji/ggab401

[17] D. Tang, and J. Hao, "A deep map transfer learning method for face recognition in an unrestricted smart city environment," Sustainable Energy Technologies and Assessments, vol. 52, no. 8, pp. 102207-102215,

2022. https://doi.org/117.176.204.213

[18] S. Langer, "Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation function," Journal of Multivariate Analysis, vol. 182, no. 3, pp. 104695-104695, 2021. https://doi.org/10.1016/j.jmva.2020.104695

[19] P. Doshi, J. Tanaka, J. Wosik, N. M. Gil, M. Bertran, S. D. Russell, and G. Sapiro, "Machine learning and video recognition for automated detection of fluid status in heart failure patients," Circulation, vol. 142, no. 3, pp. A17060-A17060, 2020. https://doi.org/10.1161/circ.142.suppl_3.17060

[20] A. Tewari, M. Zollhofer, F. Bernard, P. Garrido, H. Kim, P. Perez, and C. Theobalt, "High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 357-370, 2020. https://doi.org/10.1109/TPAMI.2018.2876842

[21] X. Y. Zhang, H. Shi, C. Li, P. Li, and P. Ren, "Weakly-supervised action localization via embedding-modeling iterative optimization," Pattern Recognition, vol. 113, no. 6, pp. 107831-107853, 2021. https://doi.org/10.1016/j.patcog.2021.107831

[22] A. B. Deshmukh, and N. U. Rani, "Optimization-driven kernel and deep convolutional neural network for multi-view face video super resolution," International Journal of Digital Crime and Forensics, vol. 12, no. 3, pp. 77-95, 2020. https://doi.org/10.4018/IJDCF.2020070106

[23] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional module for temporal action localization in videos," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 6209-6223, 2021. https://doi.org/10.48550/arXiv.2112.00302

[24] G. Yuan, J. Jinlin, and C. Wei, "Fast bilateral complementary network for deep learning compressed sensing image reconstruction," IET Image Processing, vol. 16, no. 13, pp. 3485-3498, 2022. https://doi.org/10.1049/ipr2.12545

[25] H. O. Ahmad, and S. U. Umar, "Sentiment analysis of financial textual data using machine learning and deep learning models," Informatica, vol. 47, no. 5, 2023. https://doi.org/10.31449/inf.v47i5.4673

[26] H. Zhang, J. Zhang, G. Nie, J. Hu, and W. J. C. Zhang, "Residual memory inference network for regression tracking with weighted gradient harmonized loss," Information Sciences, vol. 597, no. 1, pp. 105-124, 2022. https://doi.org/10.1016/j.ins.2022.03.047

[27] L. Cao, J. He, L. Gao, Y. Zhong, X. Hu, and Z. Li, "LWIR hyperspectral image classification based on a temperature-emissivity residual network and conditional random field model," International Journal of Remote Sensing, vol. 43, no. 9, pp. 621-645, 2022. https://doi.org/10.1080/01431161.2022.2105667

[28] R. Yamakuni, H. Sekino, M. Saito, T. Kakamu, K. Takahashi, J. Hara, H. Suenaga, S. Ishii, K. Fukushima, and H. Ito, "Prediction of anemia from cerebral venous sinus attenuation on deep-learning reconstructed brain computed tomography images," Journal of Computer Assisted Tomography, vol. 47, no. 5, pp. 796-805, 2023. https://doi.org/10.1097/RCT.0000000000001479

[29] A. Kumar, S. Majee, and S. Jain, "CDM: A coupled deformable model for image segmentation with speckle noise and severe intensity inhomogeneity," Chaos, Solitons & Fractals, vol. 173, no. 3, pp. 104385-104396, 2023. https://doi.org/117.176.204.213

[30] K. Bhosle, and V. Musande, "Evaluation of deep learning CNN model for recognition of devanagari digit," Artificial Intelligence Applications, vol. 1, no. 2, pp. 114-118, 2023. https://doi.org/10.47852/bonviewAIA3202441