Multimedia Tools and Applications (2022) 81:9875–9896
https://doi.org/10.1007/s11042-022-11937-w

Deep learning network model based on fusion of spatiotemporal features for action recognition

Ge Yang [1,2,3] • Wu-xing Zou [1,2]

Received: 3 June 2021 / Revised: 30 August 2021 / Accepted: 3 January 2022 /
Published online: 14 February 2022

Abstract
In view of the problem that the current deep learning network does not fully extract and fuse spatio-temporal information in the action recognition task, resulting in low recognition accuracy, this paper proposes a deep learning network model based on fusion of spatio-temporal features (FSTFN). Through two networks composed of CNN (Convolutional Neural Networks) and LSTM (Long Short-Term Memory), the time and space information are extracted and fused; multi-segment input is used to process large-scale video frame information to solve the problem of long-term dependence and improve the prediction accuracy; The attention mechanism improves the weight of visual subjects in the network. The experimental verification on the UCF101 (University of Central Florida 101) data set shows that the prediction accuracy of the proposed FSFTN on the UCF101 data set is 92.7%, 4.7% higher than that of Two-stream, which verifies the effectiveness of the network model.

Keywords Action recognition · Deep learning · CNN · LSTM

1 Introduction

With the development of hardware technology, especially graphics processor technology in recent years, the parallel computing capability of computers has been greatly improved. Deep learning-based technologies that rely heavily on computing power show great advantages in

✉ Ge Yang
yangge@pkusz.edu.cn

[1] Key Laboratory of Intelligent Multimedia Technology, Beijing Normal University, Zhuhai 519087, China

[2] Advanced Institute of Natural Sciences, Beijing Normal University at Zhuhai, Zhuhai 519087, China

[3] Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School, Peking University, Shenzhen 518055, China

the field of computer vision. With the popularity of various video shooting devices, the number of video generations of the network has grown rapidly, and the need to automatically analyze these video contents is becoming increasingly urgent. The development of behavior recognition technology will effectively promote the development of smart security, smart campus, video content search and detection and other fields. At present, the video content behavior recognition technology has gradually shifted from the traditional manual feature selection of the deep learning network model. Unlike manual feature selection, deep learning network models are mostly end-to-end models, without manually selecting image features, but using Convolutional Neural Networks (CNN), Long Short-Term Memory Network [-33](LSTM), etc., learn the data set, and classify the video after learning the network parameters.

Judging from the domestic and foreign research and development of video behavior recognition in recent years, video behavior recognition based on deep learning has become the mainstream, and its core is to better learn and adapt to various types of behavior through appropriate network models. To improve the accuracy of video behavior recognition, it is necessary to make full use of the time and space information hidden in the video.

There are still some problems with the current research: due to the persistence of behaviors, accurate identification of a behavior often relies on a long-range segment. For example, climbing poles and pull-ups have similar sections that use both arms to lift the body. If there is no long-range analysis, the climbing poles will be mistakenly recognized as pull-ups. Obviously, a longer range of fragment analysis can improve the recognition accuracy. Existing research is not enough to extract the dynamic time feature of video, and it is generally obtained through Optical Flow and Recurrent Neural Network (RNN). However, the single feature of optical flow is not enough to fully exploit the dynamic features of video in the time dimension. The extraction of spatial features is also inadequate. The background of some behavioral video samples is similar in large area. The difference is only the subtle difference of the main body of the behavior. At this time, the existing network models are easily confused and produce false recognition. It is also a research direction to further explore spatial features and reduce the impact of unimportant information on recognition results.

2 Related work

The current video content behavior recognition technology is mainly divided into two major directions: the traditional method of manually determining features and the method of using deep learning to establish an end-to-end prediction network model.

2.1 Behavior recognition algorithm based on traditional feature extraction

The behavior recognition algorithm based on traditional feature extraction first extracts relevant visual features, and then encodes the features. After encoding, the relevant classification techniques in statistical machine learning are used to obtain the predicted classification results. The visual features are divided into local and global features. Extracting local features refers to not segmenting the foreground and background of the video, but directly locating the desired key points of interest in the video frame and extracting features on them. The extraction of points of interest includes 3D Hessian space-time key points [3], Harris points [14], Cuboid features [31] and other algorithms. However, as the name implies, the global features start from the global perspective, first segment the main body and background in the video frame, and then describe the main body area of the image. The skeleton [25], outline [8], shape [26], etc. including the visual subject are often used as global visual features. The most commonly used visual feature extraction algorithms for behavior recognition are: HOG [11](Histogram of Oriented Gradient), HOF [21](Histogram of Flow), MBH [6](Motion Boundary Histograms), etc. After extracting the relevant local and global features of the video, it needs to be encoded, such as BOVW [20](Bag Of Visual Words), Fisher Vector [3] and other solutions, and finally a SVM(Support Vector Machine) or softmax and other statistical learning methods to obtain prediction results of behavior recognition. Algorithms based on traditional manual feature extraction often have better results for specific types of videos, but feature extraction is too dependent on manual selection, and the generalization performance is poor [13].

2.2 Deep learning network mode

Unlike manual selection of visual features, deep learning network models are mostly end-to-end models, that is, there is no need to manually select image features. Instead, CNN, RNN, etc. are used to learn the data set first, and the network parameters are predicted and classified. Hinton uses the deep learning network model to improve LeNet and deepen the structure of the network. It proposes AlexNet [16], which uses less computation to learn richer, more abstract, and higher-level image features. In ImageNet [4] The recognition accuracy rate of the picture classification task exceeds 10% of the runner-up. Immediately afterwards, Karpathy took the lead in introducing deep learning methods from the field of image classification tasks to the field of video behavior recognition [16], and proposed and compared the effects of several fusion methods of multiple multi-frame features. Explored and tried to use CNN to mine the spatiotemporal fusion information of the video. Tran extended the two-dimensional CNN and proposed a 3D convolutional neural network [28](Converse 3 Dimensional, C3D). C3D is realized by applying 3D kernel convolution to the video. It is a 2D-CNN (2 Dimensional-CNN), an accuracy of 85.2% is achieved on the UCF101 [24] dataset. Simonyan believes that video contains two parts of time and space. Spatial information is obtained through video image RGB (Red-Green-Blue) frames, and time information is obtained through optical flow containing inter-frame motion information [23]. From this point of view, Simonyan proposed a CNN-based two-stream [23] network model, which uses two independent networks to simultaneously obtain and merge the temporal and spatial features in the video. The spatial network performs single-frame video image perform analysis, while the time network learns and analyzes the optical flow and merges the two network outputs through post-fusion. In addition, Simonyan also explicitly raised the problem of long-term dependence that hinders the improvement of accuracy in the field of video behavior recognition. Ng combines CNN and RNN [34], uses CNN to obtain spatial features, RNN obtains temporal features, and characterizes video in both time and space dimensions. Donahue took the lead in combining CNN with LSTM and proposed LRCN [5] (Long-term Recurrent Convolutional Networks). LRCN uses CNN to obtain the spatial characteristics of video image frames in the shallow layer, and then uses this information as the input of LSTM in the order of video time, relying on the LSTM network to learn video time information. Unlike Donahue's LRCN model, Cooijmans [2] embeds each LSTM unit in CNN. Although this network model has a slightly lower accuracy than LRCN on the UCF101 dataset, it has better results in the field of motion prediction. Feichtenhofer proposed a spatiotemporal fusion network model based on dual streams [7], which improved the accuracy of UCF101 to 92.5%. Wang proposed a multi-segment spatio-temporal fusion network model based on long-range time structure combined with sparse time sampling [29]. In 2018 Chen [1], Yan [32] and others tried to learn human behavior from the skeleton diagram. Tang [27] also uses the bone method, but first uses reinforcement learning to find the most representative frames in the video frame, and then uses the graph-based neural network to capture the dependency between the joint connection points, so as to achieve the purpose of behavior recognition. When the mainstream of academia turned to 3D-CNN(3 Dimensional-CNN) and spatiotemporal fusion models, in 2019 Jiang used 2D-CNN to propose the STM (Spatio-temporal and motion) network model [12], which achieved good results.

In current research, some deep learning network models show good recognition effect, but there are still some shortcomings: first, the spatial features in the video are not fully mined, such as the recognition error is easy to occur when different behavior backgrounds are similar; second, the video behavior time span is large, and the use of time information is not enough; third, the video is not fully integrated time and space information of frequency.

In view of the problem that the current deep learning network does not fully extract and fuse spatio-temporal information about the behavior recognition task, resulting in low recognition accuracy, this paper proposes a deep learning network model based on fusion of spatio-temporal features (FSTFN). Time and space information are extracted and merged through two networks; multi-segment input is used to process a large range of video frame information; finally, a soft attention mechanism is introduced to reduce recognition errors due to highly similar backgrounds. The main contents of this paper are as follows: (1) Use two independent networks to extract the time and space information of the video and fuse the space-time information. (2) The video is segmented. Each video sample samples multiple segments and enters the network composed of CNN and LSTM. By covering the entire video sampling range, the long-term dependence problem of video behavior recognition is solved. (3) Introduce a visual attention mechanism at the end of CNN, reduce the weight of non-visual subjects in the network model, improve the influence of visual subjects in the video image frame, and make the best use of the spatial characteristics of the video. (4) Simultaneously extract optical flow as a dynamic feature and fully mine the time information of video behavior. The horizontal and vertical optical flow fields are used as the input of CNN in the time network to further explore the dynamic characteristics of video behavior analysis. (5) The experiment on the UCF101 (University of Central Florida 101) dataset is compared with other mainstream behavior recognition network models. The results show that the proposed FSFTN prediction accuracy rate is 4.7% higher than Two-stream [23]. At the same time, the effectiveness of the innovations proposed in this paper is verified by comparative experiments.

3 Structure of the model

The fusion of spatiotemporal features networks (FSTFN) proposed in this paper is predicted by the fusion of the spatial feature network and the temporal feature network. The spatial feature network is based on CNN + LSTM (LRCN) from the CNN terminal to the LSTM input. Attention mechanism is introduced to further mine the spatial features of video images; temporal feature network learns the two features of optical flow and RGB color difference. At the same time, the video data is sampled into multiple segments and input into the two networks, which not only analyzes the long enough time range, but also does not bring too much calculation.

Figure 1 is the overall framework of the network model proposed in this paper. It can be seen that for the original video data, it is first sampled into multiple segments, each segment is a continuous video frame, and some video frames are skipped between the segments to reduce the amount of calculation; After the sample is taken, it is preprocessed to obtain RGB information, horizontal and vertical optical flow frames for network input; RGB information of the video frame enters the spatial feature network to obtain the spatial feature prediction score, and optical flow input Go to the temporal feature network to fuse and obtain the temporal feature prediction score; for a piece of video, after getting the spatial and temporal feature prediction score, fuse it to become the final prediction output of the FSFTN.

3.1 Spatial feature network

Expand the single spatial feature network in FSFTN to obtain the internal structure of the spatial feature network as shown in Fig. 2. Here, taking the calculation of a single segment as an example, the sampled segment is input into CNN after the RGB frame is decomposed. Spatial features are extracted from RGB information to recognize actions. The spatial feature network is based on CNN + LSTM (LRCN). Attention mechanism is introduced between the CNN output terminal and the LSTM input. The feature map calculated by CNN enters the attention mechanism for weighting, and then enters LSTM to extract temporal features. At the end of the network, the spatial feature network prediction score is obtained through fusion.

3.2 Temporal feature network

Figure 3 shows the internal structure of the time feature network in FSFTN. The time feature network uses the optical flow in the horizontal and vertical directions of the image and the RGB color difference. This extraction process is completed in the preprocessing stage. The optical flow frame and RGB color difference of a single segment are input into CNN to obtain

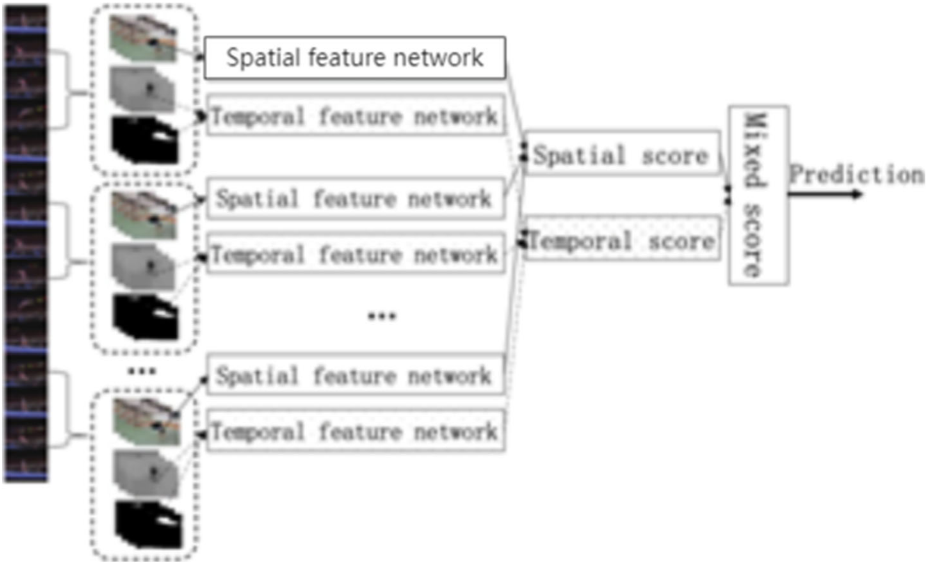


Fig. 1 Architecture of FSTFN

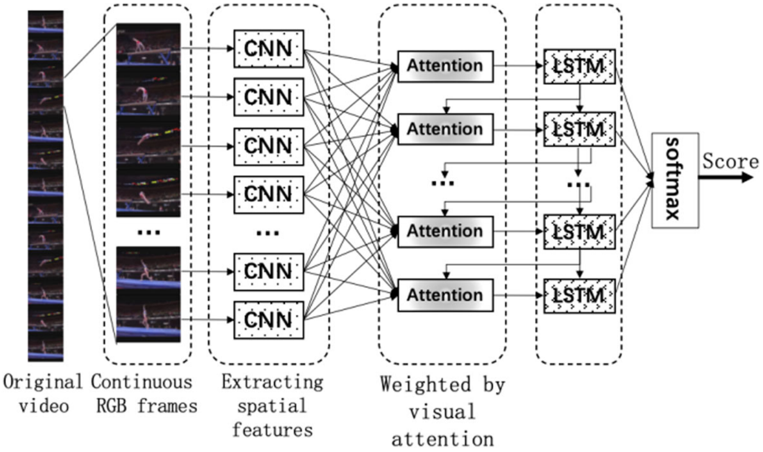


Fig. 2 Structure of spatial network

the spatial information of these dynamic features first. The feature maps obtained by CNN are weighted by the attention mechanism, and the weighted values are passed into the LSTM network to further extract their time features. The output of the LSTM network is fused to obtain the prediction score of the time network. It is worth noting here that the visual attention mechanism calculation reuses the weight of the spatial feature network calculation. In the experimental part, it is necessary to verify the contribution of optical flow and RGB color difference to the prediction results through experiments.

3.3 Preprocessing

3.3.1 Data augmentation

Due to the limited samples of the data set, and the learning ability of the network model is strong enough, the network model may have learned too many features of the data set, and

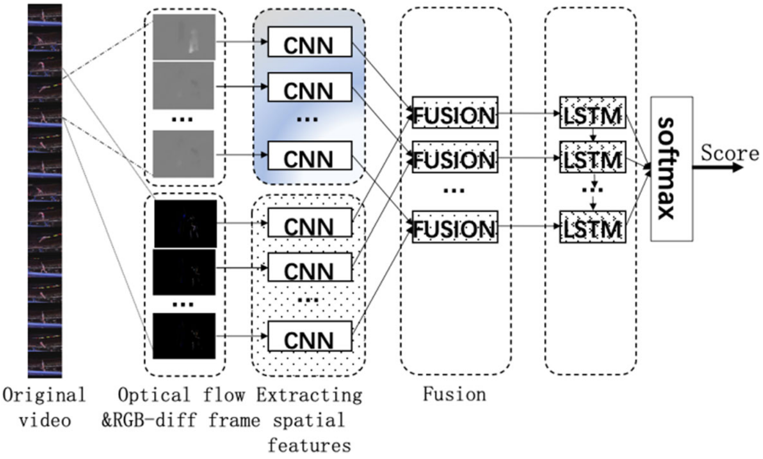


Fig. 3 Structure of temporal network

overfitting has occurred, resulting in a decline in generalization performance. However, the workload of data set sample collection and label classification is huge. It is difficult to expand the data set by collecting samples and labeling. The neural network model is an end-to-end model. If simple translation, rotation and other operations are performed on the data set, the neural network model will be regarded as a new image. This can easily expand the sample set. This paper expands the data set by three methods: horizontal mirror inversion, small angle rotation, and cropping. Small angle rotations are 30 °, 15 °, -15 ° and -30 ° clockwise. The purpose of cropping is to make the visual subject not necessarily appear in the central area of the video.

3.3.2 Segments divided

The purpose of segmentation is to learn the long-range of the video as much as possible to solve the problem of long-term dependence and improve the prediction accuracy. Since the length of the input video is not fixed, the FSFTN segmentation strategy is to divide the video into N + 1 segments first, numbered 0, 1, …, N, and continue the first half of the segment within the range of [1, N] Sampling to preserve the temporal characteristics of the video. That is, the beginning segment of the video is removed, and the remaining N segments (corresponding to numbers 1, 2, …, N) are taken, and the first $\frac{1}{2}$ intervals of the N segments are continuously sampled.

3.3.3 Extracting optical flow

The extraction of optical flow frames is carried out in the preprocessing part, and two assumptions are made in optical flow engineering calculations. First, the value of the target pixel will not change in adjacent frames; second, the adjacent pixels in the image The points have the same motion, so we can use the Lucas-Kanade dense optical flow algorithm [19], which calculates the displacement of each pixel between t and $t+\delta_t$. Assumption: The pixel $O(w,h,t)$ moves to $(w+\delta_w, h+\delta_h, t+\delta_t)$ of the next frame image after the time δ_t. Based on the first assumption:

$$O(w,h,t)=O(w+\delta_w, h+\delta_h, t+\delta_t) \tag{1}$$

Let f_w, f_h and f_t be the gradients in the horizontal, vertical, and time directions, respectively, and perform Taylor series expansion on Eq. (2) and eliminate the same terms.

$$f_w u + f_h v + f_t = 0 \tag{2}$$

Where $u=\frac{\partial w}{\partial t}$, $v=\frac{\partial h}{\partial t}$, and Eq. (2) is the optical flow equation. Further calculations need to use the second assumption that adjacent pixels have the same motion, so that 9 points in the 3 × 3pixel area have the same motion, and 9 optical flow equations can be obtained. Nine equations solve two unknowns simultaneously, which is obviously redundant, so the least squares fitting is introduced:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_j f_{w_j}^2 & \sum_j f_{w_j} f_{h_j} \\ \sum_j f_{w_j} f_{h_j} & \sum_j f_{h_j}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_j f_{w_j} f_{t_j} \\ -\sum_j f_{h_j} f_{t_j} \end{bmatrix} \tag{3}$$

By Eq. (3), the optical flow vectors at these points are obtained. Figure 4 is a visual image of the optical flow frame after optical flow pretreatment. The jittering hair dryer, fast attack fist and shaking sandbag are very obvious in the optical flow. The complex background is ignored by the whole because of stillness. It better characterizes the dynamic part of the image.

3.3.4 RGB-diff extraction

The RGB color difference is also used to measure the temporal characteristics of the image. The calculation method is relatively simple and the speed is faster than the calculation of optical flow. First, the color difference matrix $P_{diff\,t}$ obtained by differentiating the color three-channel components of the pixels corresponding to the adjacent frames P_t and P_{t-1}, for w×h pictures:

$$P_{diff\,t}[x,y] = P_t[x,y] - P_{t-1}[x,y] \tag{4}$$

Where $x \in [0, w), y \in [0, h)$. RGB color differences are obtained by superimposing P_{diff} for 5 consecutive frames. Figure 5 is the color difference extracted from the billiard behavior video, reflecting the dynamic characteristics of the video frame, which is more detailed than the dynamic characteristics of optical flow.

3.4 Attention mechanism

The attention mechanism stems from human response to external stimuli. When humans observe the outside world, they usually do not view the entire field of view as a whole (that is, each pixel has the same weight on the nervous system's influence), but is first felt The key areas of interest are attracted. It can be said that this mechanism improves the influence of the visual attention area in the human visual system, reduces the interference of useless information, and makes people have more energy to pay attention to the key content [15]. Humans glance at the scene continuously, and they will pay attention to the relevant parts of the current context. This ability is not only used to focus on the specified area, but also to interpret the same location or object in different contexts, which helps to better capture. Visual

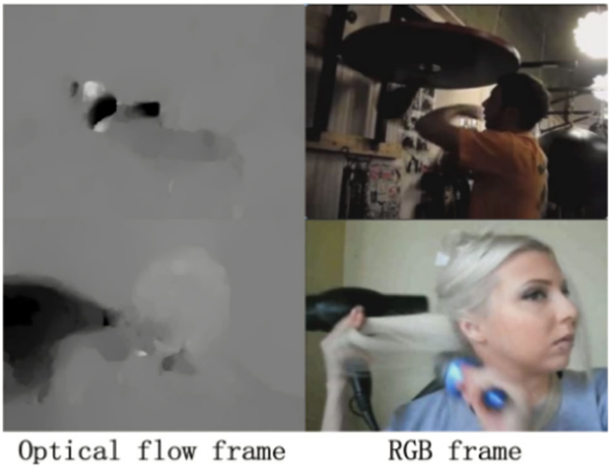


Fig. 4 Compare the optical frame with RGB frame

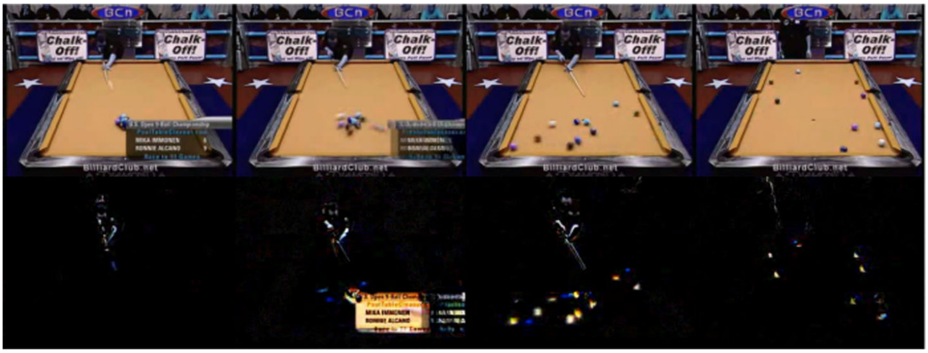


Fig. 5 Comparison of RGB-diff and RGB extracted from billiards

characteristics. The transshipment of the whole mechanism depends on two aspects: (1) Identify the key areas in the information that need attention. (2) Invest more computing resources in this area. The different levels of abstract features of the image play an important role in influencing CNN's ability to discern. LSTM also has such an explanation to some extent. The LSTM network uses the attention module to generate sequential data and update the weights of the samples in previous iterations. The attention mechanism has been introduced into CNN by researchers [18] to improve the representation form, which can identify objects from cluttered backgrounds and complex scenes.

Applying the attention mechanism to the visual neural network is to weigh the various regions of the image or feature map. For the unimportant, such as the visual background, the weight of this part of the entire network is weakened by reducing the weight; for the important, the classification predicts the key visual subject or visual focus to increase the weight.

Take the image description (Image Caption) task field with good effect of attention mechanism as an example. This kind of task can use the encoding-decoding scheme [9]. The encoding part usually uses CNN to obtain the spatial features, and the decoding uses RNN to describe the sentence in the picture shown in Fig. 6.

The attention mechanism includes a soft attention (Soft Attention) mechanism and a hard attention (Hard Attention) mechanism. The former is an attention weight assigned to each pixel of the image or feature map in the range [0,1]; 0, or 1. Soft attention is completely differentiable and can be added to the existing LSTM, because the soft attention mechanism retains the gradient and does not affect the effective propagation of the gradient during network model training. The main function of the hard attention mechanism is to take a certain area as the only

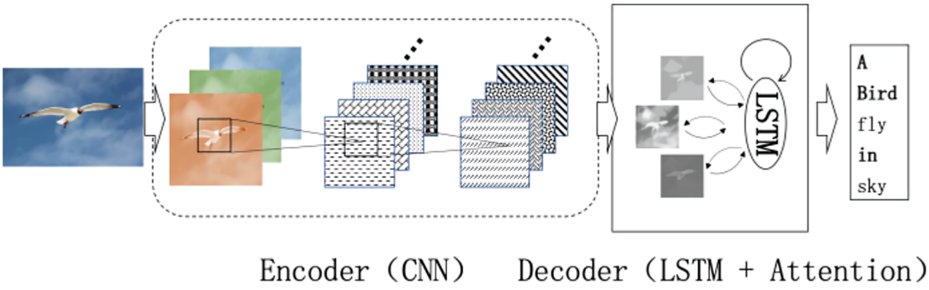


Fig. 6 Attention mechanism in image caption

area of concern, ignore other parts completely. It can be seen that in behavior recognition task scenarios, it is reasonable to apply a soft attention mechanism.

Mask calculation in attention mechanism The following describes the calculation of the mask or weight in the attention mechanism, designing the function *G*, and calculating the score s_i for each feature vector g_i^S. The calculation is based on the correlation between g_i^S and the visual subject (represented by a vector) that the attention mechanism focuses on. A higher score indicates a greater correlation. For the calculated s_i the weight γ_i is obtained via softmax, that is: $\gamma_i = softmax(s_i)$. The question then becomes how to design this scoring function *G*. In this paper, the attention scheme based on association (Concatenation-based) is adopted [30], and the image feature g_i obtained by CNN is used as the feature vector used by the attention mechanism, and *G* is a measure of the state of the hidden layer $\mathrm{h}_{\mathrm{j}-1}$ at j-1 and the function of the correlation between g_i:

$$s_{i,j} = G\left(g_i, h_{j-1}\right) = f(w_1 g_i + w_2 h_{j-1} + b) \tag{5}$$

Where $w_1 \in R^{d\times d}, w_2 \in R^{d\times d_t}, b \in R^d, s_{i,j} \in R$, and f are activation functions. Let $d_t = d, s_{i,j}$ be the score of the i-th image of the LSTM at time j, g_i represent the characteristics of the i-th image, w_1, w_2, and b are the trainable weights and biases in the scoring function *G* Set parameters. In this way, the scoring function *G* depends on $w_1$and w_2, so that the two features h_{j-1} and g_i correspond to the score $s_{i,j}$.

The reason for using the association-based scheme is that in the spatial network designed in this paper, the sampled picture frames are limited, and the hidden layer state h_{j-1} before the attention mechanism is conditional on the feature g_i of each frame output by CNN. Attention layer generates a weight $\gamma_{i,j}$, which is the mask in attention. The weight $\gamma_{i,j} \in [0, 1]$ represents the importance of the i-th frame after the calculation of the j-th attention mechanism layer, which is calculated as follows:

$$s_{i,j} = G(f_i^S, h_{j-1}) \tag{6}$$

$$\gamma_{i,j} = \frac{e^{s_{i,j}}}{\sum_{i=1}^{k} e^{s_{i,j}}} \tag{7}$$

FSTFN uses the hyperbolic tangent function tanh as the *G* function. Its bounded nature can control the output within a range; when the input deviates from 0, the curve is close to smooth, and the gradient is close to 0. The range is in $(-1, 1)$, centered on 0. h_{j-1} is the state vector of the LSTM hidden layer, and k is the number of video frames. Finally, the eigenvector Zis passed to LSTM through Eq. (8):

$$Z_j = \sum_{i=1}^{k} \gamma_{i,j} g_i \tag{8}$$

3.5 Fusion strategy

In the network model proposed in this paper, it is necessary to fuse the information of two sub-networks in time and space. For FSFTN, the methods of fusing different networks can be divided into four categories: additive fusion, splicing fusion, segment fusion and global fusion.

The specific difference of their fusion is shown in Fig. 7. The difference in fusion method will bring about differences in results, and it is necessary to determine the best fusion method through experiments.

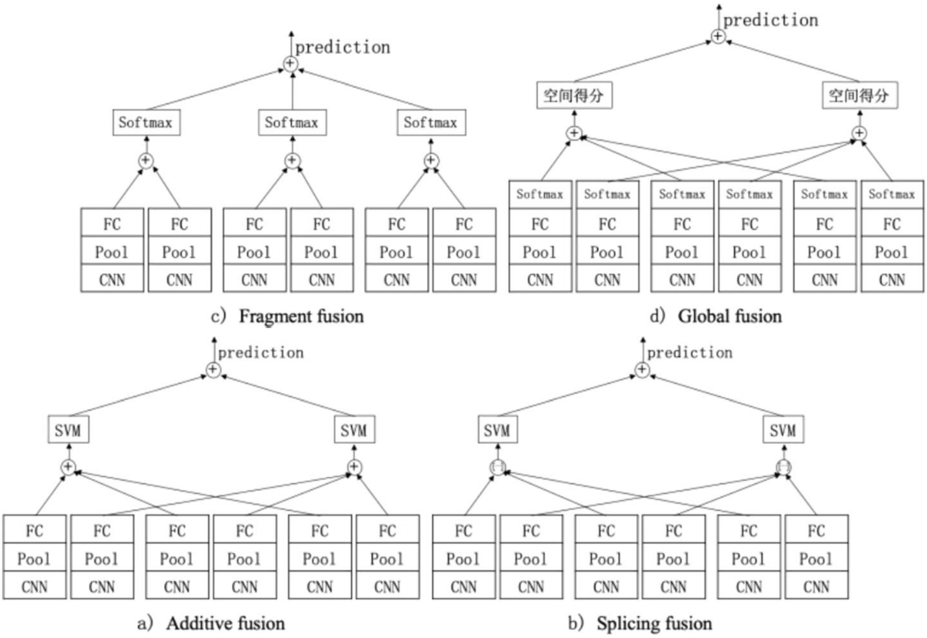
The purpose of fusion is to obtain the prediction output of the final FSFTN by using different strategies for different segment features and predictions to make up for the limitation of single feature or single segment prediction. As can be seen from the overall structure diagram of FSTFN in Fig. 1, the network model needs to consider the fusion of temporal and spatial network prediction scores.

As shown in Fig. 7, CNN represents a convolutional layer, Pool is pooling, and FC is a fully connected layer. Each video clip enters the time network and the space network separately.

(1) Additive fusion, as shown in Fig. 7(a), it means that the time network of all fragments is added after the fully connected layer to enter the time feature SVM classifier, and the spatial network of all fragments is added after the fully connected layer to enter another A spatial feature SVM classifier. During training, a global temporal feature SVM classifier and a global spatial feature SVM classifier are trained separately. Finally, the global spatial SVM and global temporal SVM are added to obtain the final prediction.

(2) Splicing fusion, as shown in Fig. 7(b), is similar to additive fusion, except that the fusion of time and space network is not additive fusion, but enters the respective SVM classifier after cascade combination, global time The SVM classifier trains a combination of all segment time network features, and the global space SVM classifier is the same.

(3) Fragment fusion, as shown in Fig. 7(c), means that the time and space network of each segment is first added and fused, and softmax classification is performed. The softmax



c) Fragment fusion

d) Global fusion

a) Additive fusion

b) Splicing fusion

Fig. 7 Four different ways of feature fusion

classification result of each segment is obtained and then weighted to obtain the prediction of the network model.

(4) Global fusion, as shown in Fig. 7(d), each network of each segment is firstly subjected to softmax to obtain the classification score, that is, twice the classification score of the number of fragments: the time network score and the space network score of each fragment. All the time network scores are added to get the time network total score, the space network total score is the same, and finally the space network total score and the time network total score are added and fused.

It can be seen that the addition fusion, splicing fusion and global are all the fusion of the temporal network features of all fragments and the spatial network features of all fragments separately to obtain the prediction after the fusion of the time network and the prediction after the fusion of the spatial network; fusion is the fusion of the temporal and spatial network features of the same segment first, and then the prediction of different segments.

When fusing optical flow features and RGB color difference features, segment fusion is used, and global fusion is used in total score prediction. For video or image classification tasks, fully end-to-end network models perform better than staged training.

4 Experiments

4.1 Experimental environment

This article is based on the CPU (Central Processing Unit) + GPU (Graphics Processing Unit) hardware environment to improve training and prediction speed. The hardware uses Intel®ฏ i7-6600 processor, NVIDIA GeForce RTX ™ 2080, 32G memory; software uses Google ’s TensorFlow deep learning framework to improve development efficiency. The prepared network model program is verified on the HMDB51 and UCF101 data sets to obtain experimental data, and compared with other network model test results. The specific software and hardware environment is shown in Table 1.

Table 1 Experimental environment

Item	Environment
CPU	Intel® Core™ i7-6600
GPU	2×NVIDIA® GeForce RTX™ 2080
Memory	2×16GB DDR4
Disk	240G SSD + 2 TB HDD
Operating system	Ubuntu 18.04
Library	OpenCV 3.4, ffmpeg 4.2, pillow 6.1.0, numpy 1.16.6, Tenserflow 1.10
IDE	Pycharm 2019.1.2
CUDA	CUDA Toolkit 10.0
Dataset	HMDB51、UCF 101

4.2 Dataset

4.2.1 UCF101

UCF101 [24] is composed of 13,320 video samples with a duration of a few seconds to a dozen seconds, 25 frames per second, and a resolution of 320 × 240, which are sorted out and marked with 101 behaviors. label. Table 2 is the distribution of UCF101 video sample frames, and Fig. 8 is its corresponding statistical histogram.

The video samples contained in UCF101 have great diversity, and the behavior is classified according to attributes as shown in Table 3. UCF101 includes different types of camera movements; various object appearances and human poses; different object ratios and viewing angles; messy backgrounds, brightness conditions, etc. The 101 behaviors in the data set are divided into 25 groups, and each group contains 4 to 7 behavior types. The behaviors of the same group have some similarities, such as similar perspectives and similar backgrounds. The behaviors in UCF101 are: the interaction between people and things; the physical behavior of people; the interaction between people; music performance; sports.

4.2.2 HMDB51

The HMDB51 (Human Motion Database 51) data set was established by the Brown University SERRE laboratory [17]. The video samples are mostly derived from YouTube and movies, which can be roughly classified into general facial movements, facial movements interacting with the outside world, and general limb movements There are 5 categories of physical movements interacting with the outside world and human interactions, a total of 51 behavior tags and 6,766 videos. Each behavior tag contains more than 102 videos with a resolution of 320 × 240.

Training and methods The FSFTN proposed in this paper was tested for two behavior recognition data sets, HMDB51 [17] and UCF101 [24]. In the HMDB51 data set, the experiment uses five-fold cross-validation. All samples S of the HMDB51 data set are randomly divided into 3 parts. Each time, 2 of the samples are used as the training set for FSFTN training, and the remaining 1 is used as the test set. To test the performance of FSFTN. This cross-validation process is repeated three times, and the results obtained from these three times are averaged to be the prediction accuracy of the final network model. On the UCF101

Table 2 Frame number distribution in UCF101

Frames	Number	Frames	Number
1-50	49	451-500	123
51-100	2297	501-550	57
101-150	3395	551-600	30
151-200	2551	601-650	18
201-250	2227	651-700	2
251-300	1515	701-750	6
301-350	526	751-800	5
351-400	284	801以上	13
401-450	222		

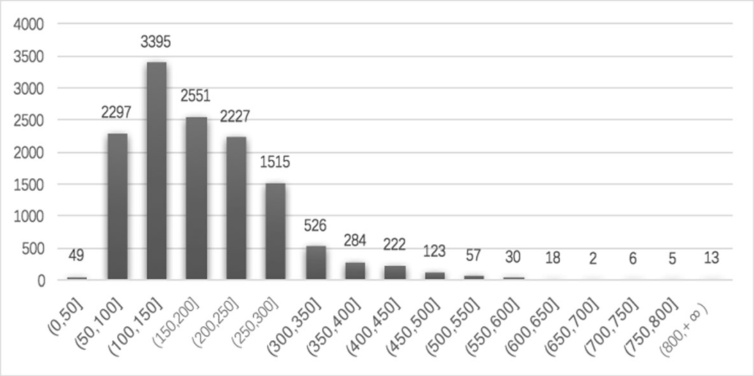


Fig. 8 Frame number distribution of video samples in UCF101 data set

dataset, due to comparison with other network models, a unified tri-fold cross-validation method is used [24], and the training of the network model uses a random batch gradient descent algorithm [35](Using the classified cross entropy loss [35] as the loss function, combined with the hardware situation, the batch of random batch gradient descent algorithm [35] is set to 128, and the momentum is 0.88. For the time feature extraction network, the learning rate is set Rate) is 0.005. In order to prevent the learning rate of oscillation when approaching the convergence point, the learning rate is reduced to 0.001 after 6000 rounds of learning; for the spatial feature extraction network, the learning rate is 0.001. After 12,000 rounds of learning, the training stopped after 4500 iterations.)to train the parameters. In terms of data enhancement, mirror reversal, small angle rotation, cropping, etc. are used. The enhanced data was segmented by the segmentation method described in 3.5, and then preprocessed as described in 3.5 to obtain optical flow and RGB color difference maps. This part of the work uses OpenCV (Open Source Computer Vision Library) + CUDA (Compute Unified Device Architecture) software framework.

Table 3 UCF101 data set behavior categories

	Attributes	Count	Attributes	Count
Body Motion	Flipping	13	Jumping Up	10
	Walking	14	Jumping Forward	13
	Running	14	Jumping Over Obstacle	3
	Riding	8	Spinning	11
	Up down	9	Climbing Up	2
	Pulling	6	Horizontal	39
	Lifting	4	Vertical Up	22
	Pushing	6	Vertical Down	26
	Diving	6	Bending	26
Number of People	One	89	Two	19
	Many	7		
Scene	Outdoor	65	Indoor	78
Posture	Sitting	33	Lying	10
	Sitting front of table	2	Handstand	7
	Standing	77		

4.3 Results and analysis

4.3.1 CNN & segments

In the experiment, when the network model is unified, AlexNet [16] and ResNet [10] are respectively selected as CNN modules to explore the influence of different CNNs on the overall prediction results. At the same time, in the multi-segment network model established in this paper, the number of segments N in 3.3.2 is a variable, and an appropriate number of segments needs to be determined. Since each additional segment requires an additional set of spatiotemporal feature extraction networks, the memory footprint and training time are greatly improved. Combined with the distribution of UCF101 video sample frame number in Table 2, take 2-6 segments to experiment separately (Table 4).

From the experimental results, it can be seen that the accuracy rate of too few segments is low, the main reason is that the small number of segments means that the spatial features are not fully utilized, that is, the long-term dependence problem mentioned in the introduction; After the value, more segment numbers cannot bring further improvement in prediction accuracy, mainly because the video context itself still has strong correlation and similarity. Obviously, the video contains a lot of redundant information, more segments are just learning this part of redundant information. Comprehensive experimental results, this article uses 4 paragraphs as the final use value.

For different CNNs, ResNet101 performs well, but AlexNet performs poorly. In the case of 6 segments, the prediction accuracy of ResNet101 on the UCF101 dataset is 28.1% higher than that of AlexNet. This may be related to the depth of the network. In the CNN comparison, the depth of ResNet101 can guarantee to learn a higher level and more abstract features, and the design of ResNet itself is also less prone to gradient disappearance. The recognition accuracy of the UCF101 data set is higher than that of HMDB51 under different CNN and different fragment numbers. This paper believes that this is related to the sample itself in the HMDB51 data set is more complicated, even the human eye has higher recognition errors.

4.3.2 Attention mechanism

Need to explore the influence of the introduction of attention mechanism on the prediction results of this network model. Design a network model of the control group and conduct a comparative experiment, where the number of segments N is 4 segments. The weighting of the attention mechanism of the control network model no longer uses the obtained mask, but

Table 4 Experimental results of different number of segments and different types of CNN were used (Accuracy)

CNN	Segments	UCF101	HMDB51
AlexNet [16]	2	49.7%	33.1%
	3	60.6%	40.3%
	4	65.1%	44.4%
	5	64.6%	44.2%
	6	64.5%	44.2%
ResNet101 [10]	2	86.7%	44.8%
	3	91.8%	54.8%
	4	92.7%	65.9%
	5	92.7%	65.9%

makes the $\gamma_{i,j}$ calculated by the soft attention mechanism in Section 3.4 fixed to a coefficient of 1 (short circuit to the attention mechanism).

From the experimental results of Table 5, it can be seen that after the introduction of the attention mechanism, the two types of CNN have achieved prediction accuracy improvements of 3.2% and 2.4% on the HMDB51 and UCF101 data sets, respectively. Accuracy. This result validates the expectation of this paper that the visual attention mechanism can increase the weight of visual subject-related features in the spatial feature extraction network and reduce the weight of non-attention areas. For example, in a video sample with a large area of similar background, it can better distinguish the subtle differences of the video subject in different behaviors.

In order to more intuitively reflect the contribution of visual attention in the network model, the weights calculated by the attention mechanism are visualized to increase the brightness of RGB images with high weights, that is, the brighter the area of interest, the darker the conversely. Figure 9 is a comparison between the RGB frame of the long jump action and the image after visualization of the attention weight. It can be seen that the long jump athlete has a small area, which is easy to ignore and leads to false recognition. However, after introducing the visual attention mechanism, the image area where the athlete is located is attracted by attention and has a higher weight to increase the proportion of the area calculated in the next step.

4.3.3 Temporal character

In the case of retaining the attention mechanism in the spatial feature extraction network, we experimented with only the spatial feature extraction network, there is a temporal feature extraction network but the temporal network only extracts optical flow features, and there is a temporal feature extraction network, which includes optical flow and RGB color The difference between the two features merges three sets of network models. The method of the control group is to set the features to be removed at the time of fusion to 0 (for adding fusion, which is equivalent to the disconnection operation), the experimental results are shown in Table 6:

It can be seen from the experimental results that no matter which data set is used to extract spatial features, the recognition accuracy is poor. The use of ResNet101 is only 79.0% on the UCF101 dataset and only 56.6% on the HMDB51 dataset. After adding the optical flow features, the recognition accuracy has been greatly improved. The HMDB51 and UCF101 data sets have been improved by 9.3% and 13.9%, respectively. The rich time features included in the visible optical flow features are conducive to recognition. However, the recognition accuracy of the network model that introduces both optical flow and RGB color difference features has decreased compared to the single optical flow time feature network. On the HMDB51 and UCF101 datasets, there was a decrease of 1.7% and 1.2%, respectively, which

Table 5 Controlled experiment on the mechanisms of attention (Accuracy)

	CNN	UCF101	HMDB51
Without attention mechanism	AlexNet	62.7%	41.9%
	ResNet101	90.3%	62.7%
Attention mechanism	AlexNet	65.1%	44.4%
	ResNet101	92.7%	65.9%

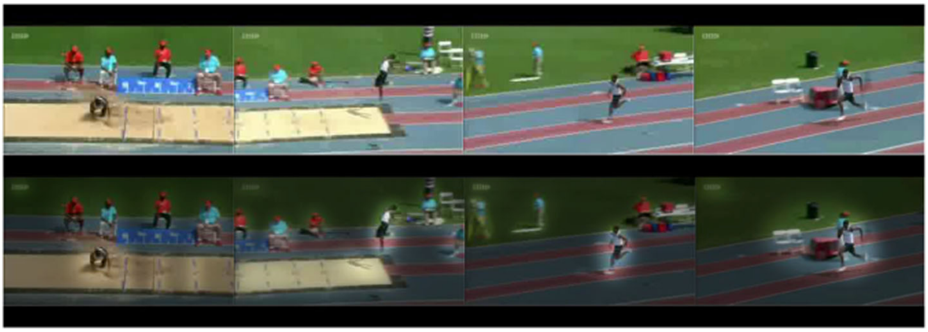


Fig. 9 Visual attention mask (weight) visualization image of long jump

shows that the addition of RGB color difference features has a negative effect on FSFTN performance.

In Fig. 10, the upper layer image is the RGB original video frame, the middle layer image is the optical flow frame, and the lower layer image is the RGB color difference. It can be seen that because the lens follows the cyclist, the characteristics of the human body are weakened in the optical flow characteristics (gray) The wall is closer to the lens and moves faster, which is strengthened in the optical flow characteristics (white); the same is the lakeside in the background, far away from the lens, the movement speed is slow, and the optical flow characteristics are weakened; the tree trunks on the roadside are. The video moves very fast and appears white. Optical flow frames are more reflective of the dynamic changes in the macro area than RGB color differences. Compared with optical flow frames, RGB color differences only reflect the edge details of dynamic differences and ignore dynamic subjects.

Figure 11 can further see this difference. The optical flow of the human body during boxing behavior reflects more of the overall moving part, while the RGB color difference only reflects the dynamic edge details.

4.3.4 Fusion strategy

In Section 3.5, four different fusion strategies are introduced. The effects of these four different fusion strategies on the results are compared through experiments. ResNet101 is used in four segments. Attention mechanism is introduced. Time network uses optical flow as input. The results are shown in Table 7.

According to the experimental results, the accuracy of FSFTN using global fusion is higher than other fusion methods. This article believes that the overall end-to-end training method for

Table 6 Control experiments with different dynamic characteristics(Accuracy)

	CNN	UCF101	HMDB51
Without temporal network	AlexNet	53.8%	34.1%
	ResNet101	79.0%	56.6%
Including optic flow feature	AlexNet	63.77%	40.3%
	ResNet101	92.7%	65.9%
Including optic flow and RGB-diff feature	AlexNet	65.1%	44.4%
	ResNet101	91.7%	64.2%

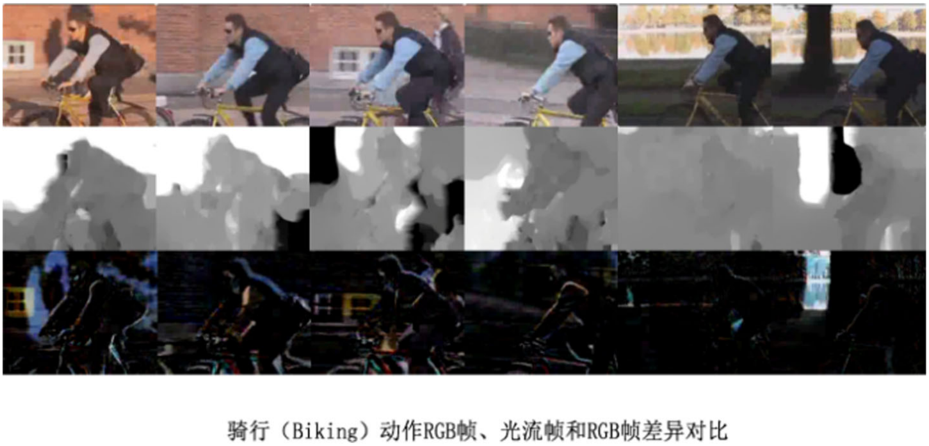


Fig. 10 Comparison of cycling behavior RGB frame, optical flow frame and RGB color difference

such large-scale classification tasks can better complement different features due to step-by-step training and adopting a global fusion strategy.

4.3.5 Accuracy and computational complexity are compared with the current mainstream network models

Light area for some videos is the most focused area of attention mechanism in Fig. 12. Dark area is the least focused area of attention mechanism. The attention mechanism improves the weight of visual subjects in the network. The attention mechanism can reduce computational load of my model in this paper.

When the number of fragments is 4, and ResNet101 is used as the CNN module, the FSFTN is compared with the experimental results of several best current behavior detection network models. FSTFN achieved an accuracy rate of 92.7% on the UCF101 dataset, surpassing other network models currently surveyed. It is 4.7% higher than the pure two-stream VGG(Two-stream Visual Geometry Group), which indicates that the introduction of the attention mechanism, the spatial and temporal fusion network based on CNN+LRCN improves the prediction accuracy on UCF101(Table 8).

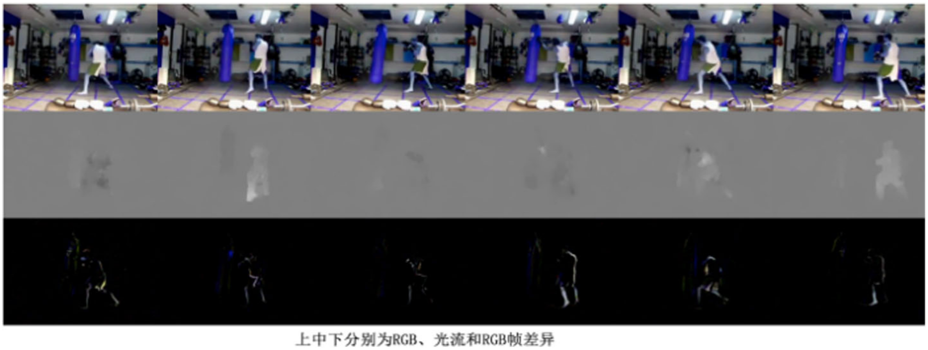


Fig. 11 RGB frame, optical frame and RGB color difference comparison

Table 7 Prediction accuracy of four fusion strategies

Fusion strategies	UCF101	HMDB51
Additive fusion	90.8%	62.9%
Splicing fusion	91.9%	63.3%
Fragment fusion	90.8%	65.7%
Global fusion	92.7%	65.9%

5 Conclusions

This paper proposes a deep learning network model FSFTN based on the fusion of spatio-temporal features to fully mine and integrate the dynamic temporal and static spatial features of the video. Specifically: (1) Use two independent networks to extract the temporal and spatial information of the video, each Each network adds LSTM on the basis of CNN to learn video time information, and fuse time and space information. At the same time, four different fusion methods were tried, and the experimental results showed that the global fusion method performed better than the other three fusion methods. (2) The video is segmented. Each video sample samples multiple segments and enters the network composed of CNN and LSTM. By covering the time range of the entire video, the long-term dependency problem of video behavior recognition is solved. In this paper, through experiments, the optimal number of segments of FSFTN on UCF101 and HMDB51 data sets is determined. (3) Introduce visual attention mechanism at the end of CNN, reduce the weight of non-visual subjects in the network model, improve the influence of visual subjects in the video image frame, and make good use of the spatial characteristics of the video. (4) Extract the optical flow as a dynamic

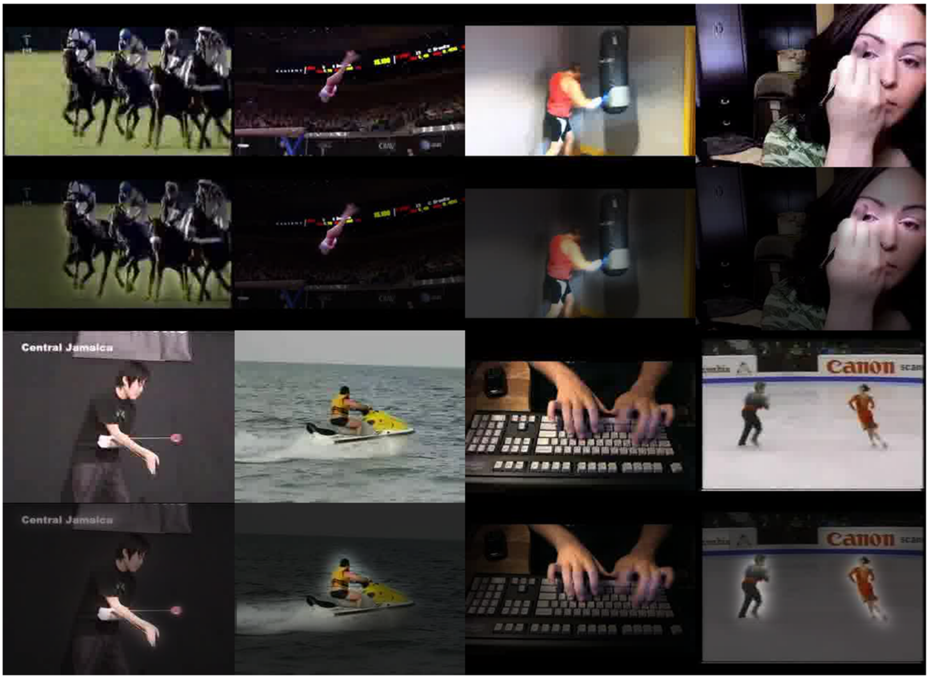


Fig. 12 Attention visualization

Table 8 Prediction accuracy and time are compared with the recognition results of mainstream network models

Network	UCF101	time(min)
ConvNets-Temporal Fusion [13]	63.9%	5.5
C3D [28]	85.2%	4.4
Two-stream VGG [23]	88.0%	5.0
LRCN [5]	82.9%	4.6
Convolutional RNNs [2]	80.7%	3.6
Snippets Fusion [34]	88.2%	4.3
Two-stream Fusion [7]	90.5%	3.8
Attention Mechanism [22]	85.0%	4.4
FSTFN(this paper)	92.7%	3.3

feature and input it into the time CNN to further mine the dynamic features of video behavior analysis. The FSFTN introducing these two features on the UCF101 data set has improved the recognition accuracy rate by 13.7% compared to when it was not introduced. In HMDB51, the accuracy rate on the data set has increased by 15.3%. The rich temporal features contained in the visible light flow are conducive to identification. Compared with the single optical flow time feature network, the recognition accuracy of the network model that introduced both optical flow and RGB color difference features decreased by 1.7% and 1.2% on the HMDB51 and UCF101 data sets, respectively. This shows that the addition of RGB color difference features FSTFN performance is counterproductive. Experiments on the UCF101 data set are compared with other mainstream behavior recognition network models. The results show that the prediction accuracy of the proposed FSFTN on the UCF101 data set is 92.7%, which is 4.7% higher than Two-stream [23]. At the same time, the effectiveness of the innovations proposed in this paper is verified by comparative experiments.

Since the optical flow calculation is performed in the pre-processing stage of the network model in this paper, it is not exactly an end-to-end model. In the future, we will study the use of a complete end-to-end model in order to achieve a good breakthrough in real-time recognition. Behavior recognition through the movement of human bones is also worthy of in-depth study. In addition, with the development of depth camera technology, visual depth, as a new feature that can be collected, can be used to better recognize video behavior, which is also the direction of future research.

Acknowledgements This research was financially supported by Major Scientific Research Project for Universities of Guangdong Province (2018KTSCX288, 2019KZDXM015, 2020ZDZX3058); Guangdong Provincial special funds Project for Discipline Construction (No.2013WYXM0122); Guangdong Provincial College Innovation and Entrepreneurship Project (S201913177027, S201913177040, 201813177028, 201813177046); Scientific Research Project of Shenzhen (JCYJ20170303140803747); Key Laboratory of Intelligent Multimedia Technology(201762005).

References

1. Chen K, Forbus KD (2017) Action recognition from skeleton data via analogical generalization[C]. Proc. 30th International Workshop on Qualitative Reasoning, 55-67
2. Cooijmans T, Ballas N, Laurent C et al (2016) Recurrent batch normalization[C]. International Conference on Learning Representations: 1-13

3. Da Silva BCG, Carvalho-Tavares J, Ferrari RJ (2019) Detecting and tracking leukocytes in intravital video microscopy using a Hessian-based spatiotemporal approach[J]. Multidimens Syst Signal Process 30(2):815–839
4. Deng J, Dong W, Socher R, Li L-J, Liand K, Fei-Fei L (2009) Imagenet: Alarge-scale hierarchical image database[C]. 2009 IEEE conference on computer vision and pattern recognition. IEEE, 248-255
5. Donahue J, Hendricks AL, Guadarrama S, et al (2015)Long-term recurrent convolutional networks for visual recognition and description[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2625-2634
6. Fan M, Han Q, Zhang X, et al (2018) Human action recognition based on dense sampling of motion boundary and histogram of motion gradient[C]. 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS). IEEE, 1033-1038
7. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 1933-1941
8. Han PY, Yee KE, Yin OS (2018) Localized temporal representation in human action recognition[C]. Proceedings of the 2018 VII International Conference on Network, Communication and Computing, 261-266
9. Hao Y, Xie J, Lin Z (2018) Image Caption via Visual Attention Switch on DenseNet[C]. 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC). IEEE, 334-338
10. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778
11. Jain A, Singh D (2019) A review on histogram of oriented gradient[J]. IITM J Manag IT 10(1):34–36
12. Jiang B, Wang MM, Gan W, et al (2019) STM: Spatio-Temporal and motion encoding for action recognition[C]. Proceedings of the IEEE International Conference on Computer Vision, 2000-2009
13. Karpathy A, Toderici G, Shetty S, et al (2014)Large-scale video classification with convolutional neural networks[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 1725-1732
14. Karthikeyan A, Pavithra S, Anu PM (2020) Detection and classification of 2D and 3D hyper spectral image using enhanced harris corner detector[J]. Scalable Comput: Pract Exp 21(1):93–100
15. Khan A, Sohail A, Zahoora U et al (2020) A survey of the recent architectures of deep convolutional neural networks[J]. Artif Intell Rev 53:5455–56516
16. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems, 1097-1105
17. Kuehne H, Jhuang H, Stiefelhagen R, et al (2013) Hmdb51: A large video database for human motion recognition[M]//High Performance Computing in Science and Engineering ‘12. Springer, Berlin, Heidelberg, 571-582
18. Li J, Liu X, Zhang M et al (2020)Spatio-temporal deformable 3D ConvNets with attention for action recognition[J]. Pattern Recognit 98:107–117
19. Meng Z, Kong X, Meng L, et al (2019)Lucas-Kanade optical flow based camera motion estimation approach[C]. 2019 International SoC Design Conference(ISOCC). IEEE, 77-78
20. Nazir S, Yousaf MH, Velastin SA (2018) Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition[J]. Comput Electr Eng 72:660–669
21. Ragupathy P, Vivekanandan P (2021) A modified fuzzy histogram of optical flow for emotion classification[J]. Journal of Ambient Intelligence and Humanized Computing 12(2):1–8
22. Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention[C]. Neural Information Processing Systems: Time Series Workshop, 1212-1225
23. Simonyan K, Zisserman A (2014)Two-stream convolutional networks for action recognition in videos[C]. Advance Neural Information Processing Systems, 568-576
24. Soomro K, Zamir A R,Shah M (2012) UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv:1212.0402:1055–1069
25. Sun B, Kong D, Wang S et al (2019) Effective human action recognition using global and local offsets of skeleton joints[J]. Multimed Tools Appl 78(5):6329–6353
26. Tanfous AB, Drira H, Amor BB (2020) Sparse coding of shape trajectories for facial expression and action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(10):2594–2607
27. Tang Y, Tian Y, Lu J, et al (2018) Deep progressive reinforcement learning for skeleton-based action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5323-5332
28. Tran D, Bourdev L, Fergus R, et al (2015) Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE international conference on computer vision, 4489-4497
29. Wang L, Xiong Y, Wang Z et al (2016) Temporal segment networks: Towards good practices for deep action recognition[C]. European conference on computer vision. Springer, Cham, pp 20–36

30. Wang X, Yu L, Ren K, et al (2017) Dynamic attention deep model for article recommendation by learning human editors' demonstration[C]. Proceedings of the 23rd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, 2051-2059
31. Yao K, Sang N, Gao C (2018) A cuboid bi-level log operator for action classification[J]. IEEE Access 6: 54147–54157
32. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Thirty-second AAAI conference on artificial intelligence, 65-77
33. Yu Y, Si X, Hu C et al (2019) A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural Comput 31(7):1235–1270
34. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al (2015) Beyond short snippets: Deep networks for video classification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 4694-4702
35. Zhang Z (2018) Improved Adamoptimizer for deep neural networks[C]. 2018 IEEE/ACM 26th International Symposiumon Quality of Service(IWQoS). IEEE, 1-2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.