Hindawi
Scientific Programming
Volume 2022, Article ID 1176279, 10 pages
https://doi.org/10.1155/2022/1176279

Research Article

Analysis of Behavioral Image Recognition of Pan-Entertainment of Contemporary College Students' Network

Hong Cui[1] **and Yuan Wang**[2]

[1]*Xi'an Technological University, Shaanxi, Xi'an 710002, China*
[2]*Xi'an Polytechnic University, Shaanxi, Xi'an 710048, China*

Correspondence should be addressed to Hong Cui; cuihong@xatu.edu.cn

Received 4 November 2021; Revised 19 November 2021; Accepted 24 November 2021; Published 17 January 2022

Academic Editor: Bai Yuan Ding

With the continuous update and iteration of network technology and technological innovation, the handheld smart media of college students will become more and more sensitive. With the advancement of economic globalization, various ideologies and cultures in the world will rapidly invade, and the "pan-entertainment" of online media may intensify. Only through the government's supervision function and the self-discipline of the internet industry, we can strictly control and screen positive values. In order to better establish the correct employment value orientation of university students and further analyze the importance of the "pan-entertainment" behavior image recognition of college students, this study analyzes the related technology and basic theory of behavior recognition. After introducing several mainstream methods, the traditional dual-stream convolutional network method is improved, and the time information and spatial information extracted by the two channels are discussed for the weighted fusion of feature maps. Finally, using R(2 + 1)D structure and dual-stream network structure design, a deep learning-based spatiotemporal convolution behavior recognition algorithm is proposed. The proposed algorithm is tested and analyzed on the datasets UCF101 and HMDB51. The specific work content is as follows: (1) to summarize the widely used video behavior classification methods proposed so far and discuss the future development. Then, it mainly analyzes the existing technical bottlenecks of some methods based on deep learning methods and summarizes and explores an efficient, stable, and accurate spatiotemporal feature joint extraction and learning method theory. (2) The design of spatiotemporal convolutional network algorithm framework is proposed, the method of segmentation processing of long video is studied, the improvement of the dual-stream network decision-level fusion method is studied, and the R(2 + 1)D network is reorganized. The network algorithm is trained and tested on the UCF-101 dataset and HMDB-51 dataset under the condition of calling the pretrained model. Finally, the accuracy is compared with the existing classic algorithms to obtain better accuracy, which proves the effectiveness of the algorithm for the "pan-entertainment" behavioral image recognition of contemporary college students.

1. Introduction

At present, the "pan-entertainment" culture is increasingly permeating every corner of people's life, and college students are most active among the consumer groups of the "pan-entertainment" culture. Due to the immature development of ideology, psychology, and other aspects and the limited ability to recognize "pan-entertainment" culture, some college students are easily affected by the negative influence of "pan-entertainment" culture, but it is difficult to identify and analyze such behavior images. With the rapid development of computer science and the modern internet world's demand for massive amounts of pictures and video information, contemporary college students' network "pan-entertainment" image behavior recognition, that is, the machine acquires the video taken by the camera, and then self-learning after preprocessing, combined with scene recognition, detect the actions of college students in the image. The machine is made smarter and more able to approach the characteristics of humans to detect images. From the 1970s to the present, some progress has been made in image recognition and analysis of abnormal motion analysis in this field. Nowadays, image behavior recognition has been successfully applied to life and can be seen

everywhere. Video recognition technology with a time dimension has become a hot and difficult point of current scientific research. For example, Wang et al. [1] adjusted and encoded DMM into pseudo-RGB images, converted their spatial and temporal behavior information into texture information, and fused three independent ConvNets networks for training and recognition. Rahmani and Mian [2] proposed a deep sequence learning view-invariant human behavior model. The method is to input each frame of a deep image into a specific convolutional neural network to learn advanced features and then transfer the human behavior in the unknown image to the model. Training and classification. Jin et al. designed a new type of RGB-D image recognition framework. The framework calculates the position deviation of 3D bone joint points and then uses the space independent nature of the joint points in the bag-of-words model to complete the vector offset and recognize human behavior. Wang et al. [3] constructed three different types of dynamic depth images, namely, dynamic depth images, dynamic depth regular images, and dynamic depth motion regular images to extract behavioral features in in-depth image sequences. Therefore, image behavior recognition technology has become one of the important contents of research and experimentation by scholars at home and abroad.

2. Behavior Image Recognition Technology Based on Deep Learning

2.1. Dual-Stream Convolutional Neural Network. Based on video processing, it can be naturally decomposed into two levels of temporal characteristics and spatial characteristics. Spatial features mainly involve the appearance description and environmental background of the subject in the "pan-entertainment" video actions of college students. Its essence is the "pan-entertainment" image recognition of static college students. The ability of spatial modeling should be strengthened to more efficiently obtain space. This deep convolutional neural network has achieved good results. The temporal feature is to clearly describe the motion feature between video frames by capturing the optical flow feature and the optical flow displacement field stacked between several consecutive frames. The network does not need to implicitly estimate motion, which reduces the difficulty of recognition to a certain extent. In 2014, Karen et al. proposed a dual-stream convolutional neural network that separately extracts and trains spatial and temporal features. Each stream of its architecture uses a convolutional neural network, and the final result is obtained after the average combination of the softmax layer scores of the two streams [4–7].

The overall framework of the dual-stream network is shown in Figure 1 below. The first network is a spatial stream network, which performs feature extraction and training on a single video frame after video preprocessing. This system is similar to image recognition. Because the action has important correlation information with the moving subject and the specific environment, the static "pan-entertainment" picture information of college students has become an important clue. The second network is the time stream, which extracts and trains the time information of the video. The optical flow diagram of multiple consecutive frames of the video is collected as input, and the motion information of the video is captured. The different information obtained between the two networks plays a complementary role [8], and it is also proved through experiments that the average result of the two convolutional neural networks is more accurate than the accuracy of the single network.

The biggest difference between dual-stream convolutional network and individual convolutional neural network training is the feature extraction of optical flow information. The optical flow feature is a set of dense optical flow, which is a set of displacement vector fields between adjacent frames. Let $d_t\ (u, v)$ be the displacement vector of the point (u, v) in the t-th frame. The optical flow is divided into two image channels, horizontal and vertical. d_t can be decomposed into a horizontal component d_x and a vertical component d_y. Frame t and frame $t+1$ are obtained together. The optical flow information obtained from a set of continuous sequence frames L is stacked to obtain a total of $2L$ channels of optical flow. Assuming that the width of the video frame is W and the height is H, then the input of the time flow network is $I_\tau = \epsilon R^{W\times H\times 2L}$ and τ represents any video frame. There are two ways to calculate I_τ. One is a very well-understood optical flow field superposition method, which is to superimpose the calculated optical flow between adjacent frames [6, 9, 10]:

$$\begin{aligned} I_\tau(u, v, 2k-1) &= d^x_{\tau+k-1}(u, v), \\ I_\tau(u, v, 2k) &= d^x_{\tau+k-1}(u, v), \\ u &= [1; w], v = [1; h], k = [1; L]. \end{aligned} \tag{1}$$

Another one is the optical flow superposition method that tracks the trajectory. Based on the inspiration of the trajectory descriptor, sampling the motion trajectory at the same position replaces the method of simple optical flow superimposition sampling at the same position in all adjacent frames. At the same time, this method also decomposes the optical flow into horizontal and vertical x and y vectors to generate $2L$ image channels. Starting from the position (u, v) of the t-th frame, let P_k be the k-th point traced along the trajectory, and then the following definitions are given:

$$\begin{aligned} I_\tau(u, v, 2k-1) &= d^x_{\tau+k-1}(P_k), \\ I_\tau(u, v, 2k) &= d^x_{\tau+k-1}(P_k), \\ u &= [1; w], v = [1; h], k = [1; L], \\ P_1 &= [u; v], P_k = P_{k-1} + d^x_{\tau+k-2}(P_{k-1}), k > 1. \end{aligned} \tag{2}$$

Both networks use the same structure and can be separately trained. This study chooses to pretrain on a larger dataset ImageNet. The spatial network stream only inputs a single RGB picture, changing the previous idea of selecting the center video frame, adopting a random selection method, and cropping the "pan-entertainment" pictures of college students to a size of 224×224. Since the optical flow network has dual vertical and horizontal channels instead of 3 RBG channels, the first

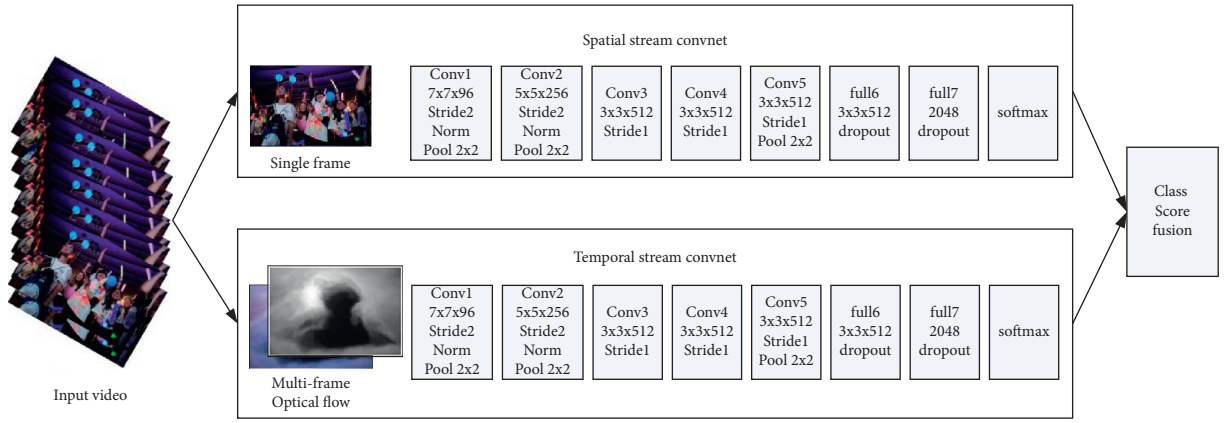


Figure 1: Dual-stream network structure diagram of video classification.

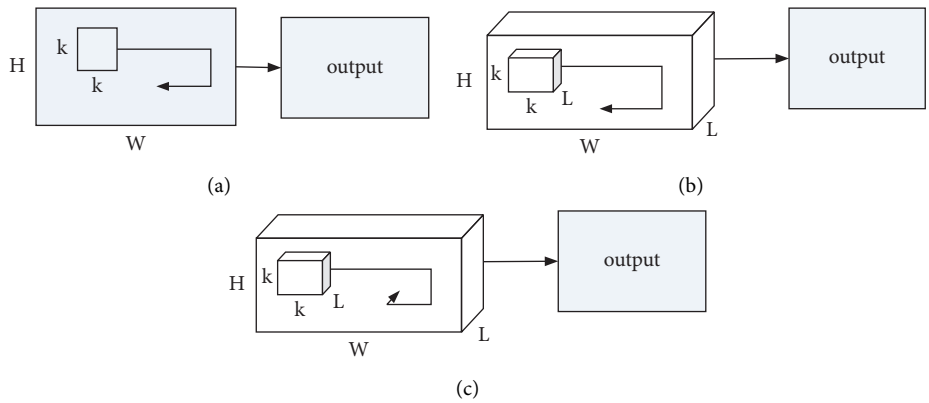


Figure 2: 2D convolution and 3D convolution.

convolutional layer is adjusted to a layer with $2L$ input channels. The previously calculated multiframe optical flow I is taken and a size of $224 \times 224 \times 2L$ from it is randomly cropped as input. After pretraining, the network was tested on UCF-101 and HMDB-51 and after multitask learning, and it was found that the method of training space and time separately is far better than the effect of single network training. At the same time, it is found through experiments that the convolutional network trained on dense optical flow still has good performance in a smaller dataset. In addition, the multitask learning used by the dual-stream network can simultaneously apply two different datasets to improve performance.

2.2. C3D Network. The theory that the dual-stream neural network separately extracts and trains temporal and spatial information opens up a new research direction for college students' "pan-entertainment" video behavior recognition, especially the extraction of optical flow features. However, the optical flow feature is also the biggest drawback of the dual-stream neural network. The optical flow vector needs to be extracted from adjacent frames, and the "pan-entertainment" video action of college students may require dozens of frames of input to accurately determine the action under special circumstances. It can be seen from this that the dual-stream network cannot extract the information of a long sequence of videos, and it is even easy to lose more other spatiotemporal information [11–13].

As shown in Figure 2, (a) is the situation obtained by 2D convolution in a static image, and the output is a two-dimensional feature map; (b) is the situation when 2D convolution is used in video operations; at this time, the image becomes a multichannel image, and the output is still two-dimensional feature map; Figure c is the operation of 3D convolution in the video, and the output is a three-dimensional feature map. The video segment is set as $c \times l \times h \times w$, where c is the number of channels, l is the length (number) of the video frame, h and w are the width and height of the video frame; the size of the convolution kernel becomes three-dimensional $d \times k \times k$, where k is the size of the convolution kernel and d is the newly added time depth; then, the output value becomes $k \times l \times h \times w$. It is inferred from this that 3D convolution better completes the spatiotemporal modeling of video data [14], and the network can directly input video data without any preprocessing.

For college students' "pan-entertainment" video processing, an effective descriptor should have the following four characteristics: first, versatility and distinguishability; second, the descriptor should be compact. There are millions of videos, and the compact descriptors in the processing process can strengthen the retrieval task and storage; third,

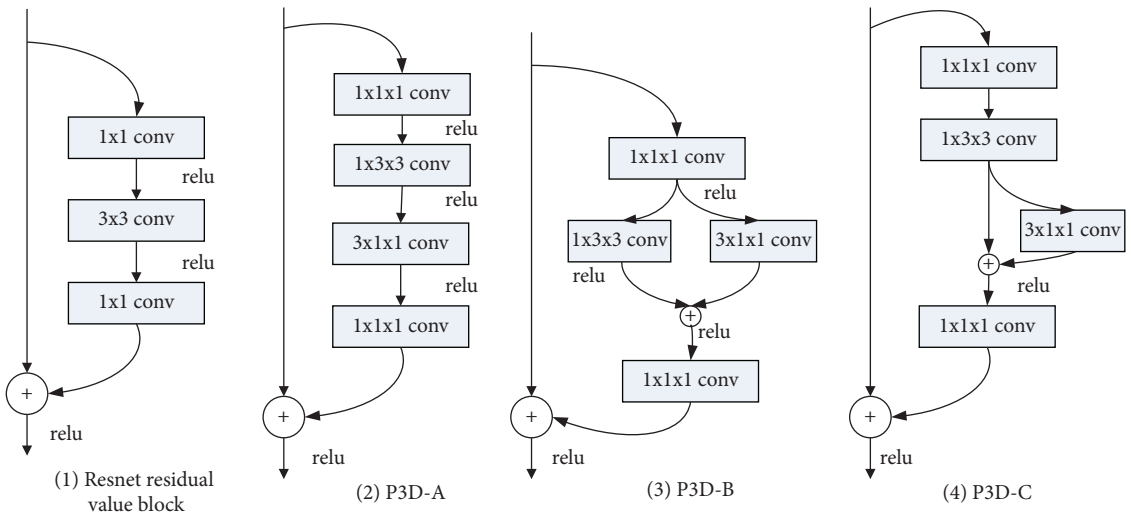
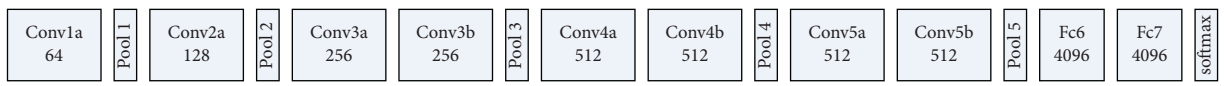


Figure 3: Three kinds of pseudo-3D block. (a) ResNet residual value block. (b) P3D-A. (c) P3D-B. (d) P3D-C.

the efficiency of calculation. Increasing the calculation speed is one of the goals we have been pursuing; fourth, simplicity. The features extracted by the 3D convolutional network contain other information such as the subject, background, and other information in the "pan-entertainment" video of college students, which can cope with various recognition requirements. The network architecture of C3D is shown in Figure 2. The size of the convolution kernel is $3 \times 3 \times 3$, and the stride is set to $1 \times 1 \times 1$. In order to keep the dimension unchanged, the size of the first pooling core is set to $1 \times 2 \times 2$, and the stride is $1 \times 2 \times 2$. The core size of the remaining pooling layer is $2 \times 2 \times 2$, and the stride is $2 \times 2 \times 2$. The network consists of 8 convolutional layers, 5 maximum pooling layers, and 2 fully connected layers and finally connected to the softmax layer. C3D chose the largest dataset Sports-1M for training. Compared with the UCF101 trained by the previous 2D convolutional network, it exceeded 5 times the category and 100 times the total number of videos. The input video is cropped into 16 112×112 segments, and the picture is horizontally flipped by 50% for data enhancement.



What are the differences in the learned features between the C3D network and the classic 2D network? This study uses the deconvolution method to detect what the 3D network has learned. After observation, it is found that C3D tends to track the subsequent development of motion information from the appearance information of the previous few frames. The C3D learning method can not only capture the appearance information of the moving subject well but also learn the variable motion information, which can be selectively learned. The proposal of the C3D network has achieved huge progress from the extraction of low-level semantics to the extraction of high-level abstract semantics, surpassing traditional deep learning methods and surpassing traditional manual methods. It is a simple and efficient model with good prospects. However, due to a large amount of calculation of the three-dimensional convolution operation, the large number of parameters generated in the C3D network has also become a problem to be solved.

2.3. R(2 + 1)D Network. After research and exploration, it is found that the 2D convolutional network is still the best choice in the analysis of action recognition. The 3D convolutional network that introduces spatial and temporal dimensional features has also brought significant progress to the research of video recognition. Unfortunately, both are flawed. The 2D convolutional network can easily lose a lot of key information due to the inability to extract long-sequence video information; the 3D convolutional network is too computationally expensive, resulting in too much speed and too many parameters, which requires a lot of storage space and other problems. In 2017, Zhaofan Qiu et al. proposed Pseudo-3D Residual Net (P3D ResNet). Based on the ResNet network, the 3D convolution kernel undergoes a series of deformations, so that the 2D and 1D convolution kernels separately operate, which not only expresses the timing information well but can also greatly reduce the amount of calculation, making the network easier to optimize. On the basis that the residual idea of the ResNet network has achieved good results in the 2D convolutional network, the 3D convolution is split into 1D convolution about time information and 2D convolution about space information. The question that needs to be discussed is what kind of connection state is between 2D and 1D convolutions. Both series and parallel will affect the final effect. The three deformation methods are shown in Figure 3 below [15].

P3D has designed three convolution deformation methods: P3D-A decomposes $3\times3\times3$ convolution integrals into $1\times3\times3$ and $3\times1\times1$ and connects them in series. The output of a 2D convolution is used as the input of 1D convolution; P3D-B does the same decomposition. At this time, 2D convolution and 1D convolution are parallel, and there is no connection between the two convolutions. The final result is determined by the sum of the two convolutions; P3D-C is the combination of the two structures P3D-B and P3D-A. It needs to be pointed out that the bottleneck design in each residual block unit in the ResNet network structure makes that there are two convolutional layers of $1\times1\times1$ before and after, which are used to reduce the size of the input dimension and restore the output dimension. In this way, the effect of reducing computational complexity is achieved. In order to evaluate these three deformation methods, based on ResNet-50, the original residual unit is directly replaced with the three designed by P3D. It is found through experiments that the P3D-A series connection method achieves the best effect. Based on the above situation, in 2018 Du Tran et al. proposed the R(2 + 1)D convolution block and conducted related experiments on the kinetic dataset to demonstrate its usability and prove that the 3D convolution kernel is decomposed into separate space and time. Separate extraction instead improves accuracy. The proposal of the R(2 + 1)D network is also based on the changes made by R3D [8] that have applied ResNet to the 3D convolutional network, using the basic residual structure, without using the bottleneck design in P3D. Each residual block consists of two convolutional layers, and each layer has a ReLU activation function. Let x denote the input of $3\times L\times H\times W$, L be the number of frames edited in the input, H and W denote the height and width of the video frame, respectively, and 3 denote the picture RGB channel. Let z be the tensor calculated by the i-th convolutional block in the residual network and the output of the i-th residual block, $\mathscr{F}(z_{i-1};\ \theta_i)$ performs two convolutional layers through the weight θi synthesis and the role of the activation function ReLU [16].

$$Z_i = Z_{i-1} + \mathscr{F}(z_{i-1};\ \theta_i). \tag{3}$$

In order to keep the R(2 + 1)D network and the R3D network roughly the same amount of parameters, the hyperparameter M_{i} is introduced.

The R(2 + 1)D convolution kernel is specifically composed of $N\ \mathrm{i}-1\times1\times d\times d\ Mi$ 2D convolution kernel and $Mi\times t\times1\times1\ Ni$ time 1D convolution kernel, instead of 3D Of $N\ \mathrm{I}\ N\ \mathrm{i}-1\times t\times d\times d$ convolution kernel, which has the following relationship.

$$\begin{aligned} N_{i-1}\times t\times d^2\times N_i &= N_{i-1}\times d^2\times M_i + M_i\times t\times N_i,\\ N_{i-1}\times t\times d^2\times N_i &= \left(N_{i-1}\times d^2 + t\times N_i\right)\times M_i,\\ M_i &= \left\lfloor \frac{td^2N_{i-1}N_i}{d^2N_{i-1} + tN_i} \right\rfloor. \end{aligned} \tag{4}$$

Figure 4 shows the difference between (2 + 1)D convolution and 3D convolution. It is supposed that the input is single-channel spatiotemporal information. On the left is 3D convolution, using a convolution kernel of size $t\times d\times d$, where t is time, and d represents the height and width of the space. On the right is the (2 + 1)D convolution block, which calculates spatial 2D convolution and time 1D convolution, respectively, and sets the number of 2D convolutions to M, so that the number of parameters in the (2 + 1)D block is the same as the entire 3D volume The number of parameters of the block matches [17, 18].

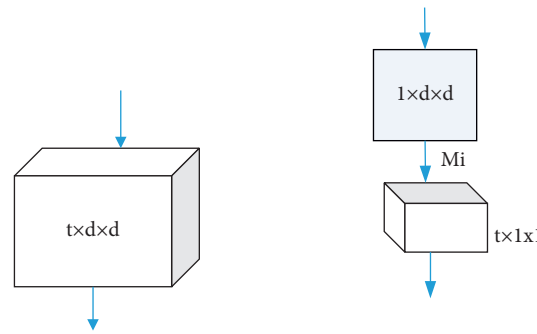


Figure 4: 3D convolution kernel and (2 + 1)D convolution kernel.

Compared with the convolution kernel of time convolution and spatial convolution, factoring into the 3D convolution kernel has two advantages. First, in the case of the same amount of parameters, the (2 + 1)D convolution kernel has doubled nonlinear mapping and has better network expression ability. Second, the (2 + 1)D network is easier to optimize, the loss rate of training and testing is lower, and the number of network layers is allowed to be deeper. This is confirmed by experiments on the large kinetic dataset. Whether it is the classic ResNet-3D network or the deformed network with the 2D convolutional layer and the 3D convolutional layer interlaced, the final result is not as good as R(2 + 1)D on the internet. One of the findings is that the effect of the ResNet-3D network, which independently separates spatial convolution from temporal convolution, is still stronger than that of hybrid 2D-3D convolutional networks. This further illustrates that separating space and time calculations is important to improve network performance factor. It is worth noting that the R(2 + 1)D network effect of the input optical flow graph still has a lot of room for improvement. Although the dense optical flow method using the Franeback algorithm improves the efficiency, the optical flow accuracy is not high, and it is worthy of in-depth study. In addition, in the experiment, the network obtained the best results after fusing the results of the input RGB image with the results of the input optical flow diagram. It can be seen that the timing characteristics and spatial characteristics are both very important.

3. "Pan-Entertainment" Behavioral Image Recognition of College Students Based on Spatiotemporal Convolutional Network

3.1. Dual-Stream Space-Time Integration Network Design. In this study, the algorithm chooses to model the spatial action and temporal information of the input relatively high-resolution "pan-entertainment" images of college students in the shallow network and uses dual-channel 2D Conv to

extract temporal and spatial features in parallel. In the later stage of the deep network part, 3D Conv is again used to perform space-time modeling. The spatiotemporal-r(2 + 1)d end-to-end model proposed in this study performs weight adjustment and fusion of the features extracted by the dual-stream neural network to obtain the middle-level semantic features, which are input into the R(2 + 1)D network for further learning and complete behavior recognition. At the same time, in order to be effective for long-term video, the video segmentation method is used in the input part of this study. The overall framework is shown in Figure 5.

The network framework is divided into three modules, namely, segmentation and RBG image preprocessing and optical flow image calculation of the input "pan-entertainment" video of college students, a weighted fusion of the dual-stream network part, and respatial modeling of the R(2 + 1)D network. First, the T_1, T_2, T_3, …, Tn video is divided into K segments of equal length $\{s_1, s_2, s_3, \ldots, s_k\}$ where S_0 consists of multiple frames. Each video frame T_n is randomly sampled from S_0 and will be used as the input of the spatial network. The extracted "pan-entertainment" feature map is x_a, where a represents the "pan-entertainment" feature map extracted from the spatial domain. The time domain network corresponds to the input continuous optical flow image and the set t time as the corresponding time of the video frame T n; then, the position of the L continuous optical flow frame pictures corresponds to time t in the time domain. The time domain feature map obtained is x_b, where b represents the "pan-entertainment" feature map extracted from the time domain network. Then, the weighted sum fusion method is used to obtain consecutive spatiotemporal feature maps that are subsequently input to the R(2 + 1)D network.

$$M \in R^{H\times W\times D\times\lambda}\left(\lambda = \theta x^{ak} + \mu x^{bk}\right). \tag{5}$$

3.1.1. Data Preprocessing of Input Airspace Network. The spatial network takes RBG pictures of static video frames as input. In the training process, the VGG-M-2048 model with pool1 and pool2 is selected, and the fully connected layer is removed for later feature fusion. This study deals with video segmentation. Among K segments of equal length, each segment randomly selects 1 frame of the image and crops it into 224 × 224 size. In order to enhance the generalization ability of the network, training samples are increased to prevent overfitting of the deep learning model, and a series of data enhancement operations are performed on the input images. This study adopts three processing methods for the input single-frame image:

(i) the picture is horizontally flipped, (ii) the angle of the picture is rotated, and (iii) the horizontal and vertical offset and shift transformation are carried out on the picture. After data expansion of the image data, the sample is increased by multiples to prevent overfitting to a certain extent.

3.1.2. Data Preprocessing of Input Time Domain Network. In the time domain network input part of the dual-stream network, this study uses the TV-L1 method to calculate the optical flow, which is divided into horizontal and vertical optical flow diagrams. The optical flow feature is a set of dense optical flow, which is a set of displacement vector fields between adjacent frames, which can be used to extract the "pan-entertainment" information of college students and play an important role in video recognition. The optical flow image imported into the input of the network framework designed in this study contains the motion information of each static video frame image, which improves the correlation of spatiotemporal features on pixels and the robustness of processing video frame sampling. As shown in Figure 6, the optical flow diagram is a grayscale image calculated by decomposing the optical flow data into horizontal and hammer direction vectors, and there are 2L image channels.

In this study, $L = 10$ is set, and the horizontal and vertical optical flows of 10 consecutive frames are stacked to form 20 dense optical flow images as the input of the time domain network.

This study uses the VGG-M-2048 model pretrained on the ImageNet dataset. However, the RGB image with the number of channels as 3 in the first conv1 does not match the input data of the time domain network. For this problem, we use the cross-modality cross pretraining method. The weights of conv1 are averaged and copied into 20 copies as the weights of the time domain network conv1. Other weights remain the same to achieve the time domain network pretraining network parameter matching.

3.2. Improved Design of R(2 + 1)D Structure. The spatiotemporal-r(2 + 1)d end-to-end model is the spatiotemporal feature map obtained after the feature fusion of the dual-stream network. Here, the spatiotemporal dual-stream network removes the fully connected layer because the output of the fully connected layer is high-level semantic features that will affect the image. The information on the time axis is not conducive to subsequent modeling. Then, the obtained spatiotemporal fusion feature map is input into the R(2 + 1)D network, which has a better effect than the C3D network. R(2 + 1)D uses the classic network ResNet. In the case of the fusion and series connection of the dual-stream network and the 3D convolutional network, the formed deep network is prone to the disappearance of the gradient, which makes the network effect worse. By inserting jump connections, ResNet directly propagates the gradient from the lossy layer at the end of the network to the early layer close to the input, which is to simplify the training structure of the deep structure. ResNet can solve the problem of gradient disappearance in deep networks to a certain extent.

Since the output part of the dual-stream network is a 28 × 28 feature map, this study removes the first three convolution blocks on R(2 + 1)D-34 and R(2 + 1)D-50, conv1_1, and conv2_1. Spatiotemporal downsampling is performed on the top and is implemented with a convolution step size of 2 × 2 × 2. Finally, a 512-dimensional feature vector is an output, and the final result is output through the fully connected layer and the softmax layer. M is the number of 2D space convolution kernels after 3D convolution kernels are decomposed into (2 + 1)D

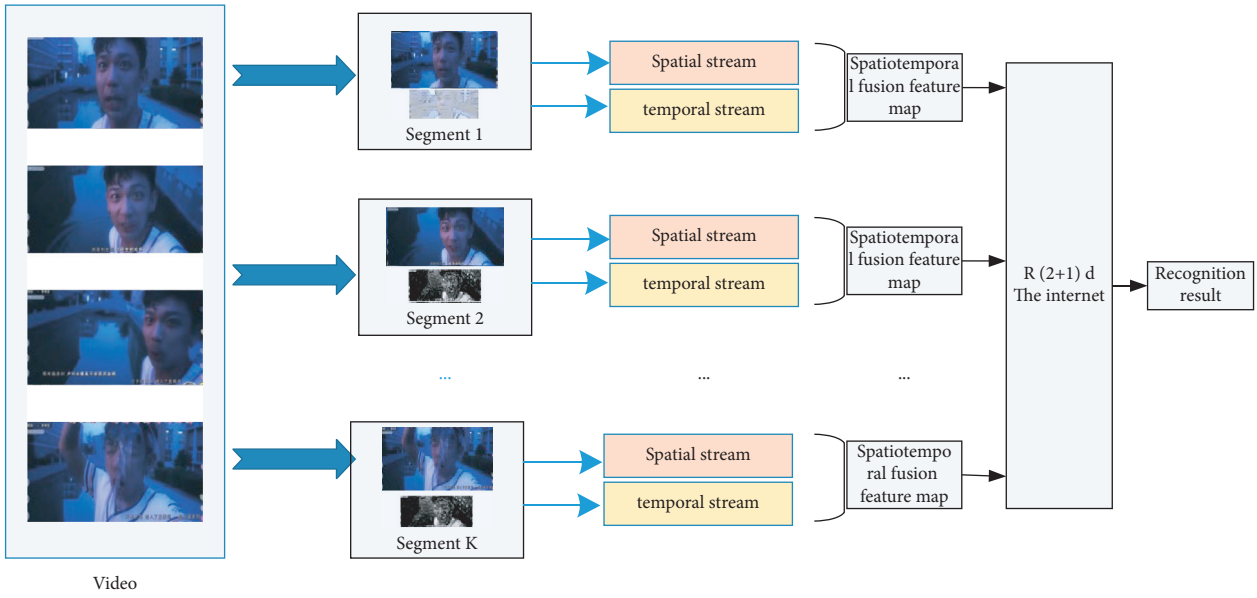


Figure 5: Spatiotemporal-r (2 + 1) d architecture design.

Figure 6: Optical flow diagram of continuous video frames.

convolution kernels. The network structure is shown in Table 1 below.

4. Experiment and Result Analysis

4.1. Experimental Data Set. In this study, UCF-101 and HMDB-51 are selected for evaluation experiments based on the most widely used datasets in the field of deep learning video behavior recognition. Next, we will introduce the HMDB-51 dataset. The HMDB-51 dataset comes from YouTube and Google videos collected from the internet. The pixels of the video frame are 320 × 240, and the average duration is 3.0 seconds. "Pan-entertainment" movements are divided into five types: (1) general facial movements; (2) facial movements with object manipulation; (3) whole-body movements; (4) body movements that interact with objects; and (5) body movements caused by human interaction. Compared with the UCF-101 dataset, the dataset comes from real scene videos, and the background greatly changes. The small amount of data results in the limited training of the network and the existing algorithms generally have low recognition rates for it, making it one of the current challenging datasets. Figure 7 shows an example of some actions of HMDB-51.

Both UCF-101 and HMDB-51 have a standard three-segment evaluation protocol. This study evaluates the network performance based on the average recognition accuracy of all splits. This study divides the two datasets into the training set and test set in the experiment.

4.2. Parameter Setting. The dual-stream network module uses the VGG-M-2048 model pretrained on the ImageNet dataset, and the R(2 + 1)D network module uses the R(2 + 1)

Table 1: Improved network structure based on R(2 + 1)D-34/R(2 + 1)D-50.

Number of layers	Output size	R (2 + 1) D-34	R (2 + 1) D-50
Conv1_x	$\lambda \times 14 \times 14$	$\begin{bmatrix} \begin{pmatrix} 1\times3\times3 & M_1 \\ 3\times1\times1 & 256 \end{pmatrix} \\ \begin{pmatrix} 1\times3\times3 & M_1 \\ 3\times1\times1 & 256 \end{pmatrix} \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1\times1 & 256 \\ \begin{pmatrix} 1\times3\times3 & M_2 \\ 3\times1\times1 & 256 \end{pmatrix} \\ 1\times1\times1 & 1024 \end{bmatrix} \times 6$
Conv2_x	$\lambda/2 \times 7 \times 7$	$\begin{bmatrix} \begin{pmatrix} 1\times3\times3 & M_2 \\ 3\times1\times1 & 512 \end{pmatrix} \\ \begin{pmatrix} 1\times3\times3 & M_2 \\ 3\times1\times1 & 512 \end{pmatrix} \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1\times1 & 256 \\ \begin{pmatrix} 1\times3\times3 & M_4 \\ 3\times1\times1 & 256 \end{pmatrix} \\ 1\times1\times1 & 2048 \end{bmatrix} \times 3$
	$1 \times 1 \times 1$	Spatiotemporal pooling, fc layer with softmax	

Figure 7: Some examples of "pan-entertainment" actions of HMDB-51.

D-34 and R(2 + 1) D-50 model. Using small batch stochastic gradient descent (SGD), the loss function is cross-entropy loss, and the BN (batch normalization) layer is added to the network to accelerate the network convergence speed, to a large extent, to prevent overfitting and to improve the problem of gradient disappearance. Tables 2 and 3, respectively, list the detailed parameters on the two datasets.

The network is trained on two datasets. The input of the spatial network is an RGB image cropped to 224×224; the input of the time domain network is 20 optical flow images of 10 consecutive video frames, the size of which is also 224×224. On UCF-101, the batch size of the spatiotemporal-r(2 + 1)d-34 dataset is 128, the initial learning rate is set to 0.001, and the learning rate decays to 1/10 of the original for 5,000 iterations, with a total of 15,000 iterations. The batch size of the spatiotemporal-r(2 + 1)d-50 dataset is 64, the initial learning rate is 0.001, and the learning rate decays to 1/10 of the original for 30,000 iterations per 10,000 iterations. The batch size of the spatiotemporal-r(2 + 1)d-34/50 dataset on HMDB-51 is set to 64, the initial learning rate is 0.001, and the learning rate decays to 1/10 of the original every 3,000 times, with a total of 10,000 iterations. The network is based on the transfer learning method, using a pretraining model for training, the initial learning rate is set to a smaller value, and then, the weight of each network is fine-tuned.

4.3. Comparison of Fusion Feature Weights. The sum of the fusion method of spatial convolutional network and time convolutional network forms a fusion method of different weights by setting the spatial weighting coefficient θ and the time domain weighting coefficient μ, and the final output feature map is obtained by adding the weights. The fifth convolutional layer of the dual-stream convolutional network module is chosen to experiment with different strategies on the fusion coefficient ratio. Based on the experiment, take the video segmentation into 3 groups and compare the average accuracy of 3 groups on the UCF-101 dataset and HMDB-51 dataset (all splits). At the same time, in order to compare the performance of the 34th layer and 50th layer of the improved r(2 + 1)d network structure, we use r(2 + 1)d-34/50 to separately experiment to discuss the weight ratio of the dual-stream fusion method. The results are shown in Tables 4 and 5. The weight ratios are taken in 7 different proportions. It can be seen that when the spatial feature map accounts for a large proportion, the recognition accuracy decreases; conversely, when the temporal feature map accounts for a larger proportion, the accuracy increases. It can be concluded that the time information extracted by the time domain network plays an important role in the overall network performance. In summary, it is found that whether it is using r(2 + 1)d-34 or r(2 + 1)d-50, θ: $\mu = 4:6$, the network recognition performance is the best, reaching the highest accuracy.

4.4. Overall Network Performance Evaluation. In order to prove that the network framework proposed in this study has certain advantages, it is compared with some existing classic algorithms on the public datasets UCF-101 and HMDB-51. Table 6 lists different comparison algorithms, including traditional hand-designed feature algorithms (IDT) and algorithms based on deep learning. It can be clearly seen that the accuracy of the framework based on deep learning algorithms (no. 3, 4, and 5) is greatly improved compared to the different feature encoding methods (no. 1 and 2) based on dense trajectories. Among them, the C3D algorithm has a relatively poor effect due to too many network parameters. However, the 3D Conv proposed by the C3D network opens the research direction of spatiotemporal convolutional neural networks based on the 3D convolution kernel, which is of great significance for video behavior recognition. At the same time, it can be seen that the accuracy of the most primitive dual-stream convolutional neural network algorithm has been improved after the LSTM recurrent neural network is added, indicating that a reasonable combination of the dual-stream network with other methods can improve the recognition effect. Analyzing the latest research results in recent years, the P3D network

Table 2: Parameter settings on UCF-101 dataset.

The Internet	cpoch	Batch_size	fropout	Initial learning rate	Momentum
Spatiotemporal-r(2 + 1)d-34	200	128	0.5	0.001	0.9
Spatiotemporal-r(2 + 1)d-50	200	64	0.5	0.001	0.9

Table 3: Parameter settings on HMDB-51 dataset.

The internet	cpoch	Batch_size	fropout	Initial learning rate	Momentum
Spatiotemporal-r(2 + 1)d-34	180	64	0.5	0.001	0.9
Spatiotemporal-r(2 + 1)d-50	180	64	0.5	0.001	0.9

Table 4: The average accuracy of r(2 + 1)d-34 fusion ratio of different spatiotemporal feature maps%.

$\theta_{\text{spatial}} : \mu_{\text{temporal}}$	UCF-101	HMDB-51
2 : 8	86.4	61.1
3 : 7	87.6	61.8
4 : 6	88.3	62.0
5 : 5	86.2	61.7
6 : 4	85.0	60.8
7 : 3	84.4	60.3
8 : 2	83.8	60.5

Table 5: The average accuracy of r(2 + 1)d-50 fusion ratio of different spatiotemporal feature maps%.

$\theta_{\text{spatial}} : \mu_{\text{temporal}}$	UCF-101	HMDB-51
2 : 8	90.5	65.3
3 : 7	90.7	65.7
4 : 6	92.1	66.1
5 : 5	91.8	65.8
6 : 4	90.4	64.2
7 : 3	89.2	62.5
8 : 2	88.5	59.8

Table 6: Comparison of accuracy rate of recognition of "pan-entertainment" image and video behavior of college students with different algorithms%.

Serial number	Method	UCF-101	HMDB-51
1	IDT + FV	85.9	59.1
2	IDT + bovw	87.9	60.9
3	Dual-stream(VGG-M)	88.0	59.4
4	Dual-stream+(LSTM)	88.6	—
5	TDD + IDT	91.5	65.9
6	C3D	85.2	—
7	P3D ResNet	88.6	—
8	MiCT-Net	88.9	63.8
9	Spatiotemporal-r(2 + 1)d	92.1	66.1

model using the ResNet network improved based on the traditional C3D network is an initial attempt to factorize the 3D convolution kernel. Compared with the traditional C3D, the effect obtained on the dataset UCF-101 greatly improved. Proposing a hybrid 2D and 3D convolutional deep network MiCT-Net compared to the traditional single 2D convolutional network and 3D convolutional network, the accuracy of the two datasets is also improved.

This study is based on the improvement of the dual-stream convolutional network algorithm. For long-term video spatiotemporal modeling, the recognition accuracy of 92.1% and 66.1% is achieved on UCF-101 and HMDB-51, respectively. Compared with the dual-stream method, the improvement is 4.1% and 6.7%, respectively. Compared with other classic algorithms, the method in this study also obtains higher accuracy. At the same time, the algorithm in this study also realizes the end-to-end network structure and realizes the effectiveness of the recognition task based on the "pan-entertainment" video behavior of college students.

5. Conclusions

This study mainly explores the recognition of "pan-entertainment" image and video behavior of college students based on the spatiotemporal convolutional neural network. The research work is as follows:

(1) The methods of "pan-entertainment" video behaviors proposed by college students so far are summarized and discussed, and the future development is discussed. By analyzing the technical bottlenecks of the current network structure in deep learning, an efficient, stable, and accurate method is summarized and explored.

(2) The dual-stream neural network is improved, and the classification decision fusion method is changed. The time information and spatial information separately extracted in the dual-channel 2D Conv are combined with "pan-entertainment" feature maps to form spatiotemporal feature maps and then behavior classification. The specific fusion method and fusion location are discussed, the influence of fusion location on network learning is explained through experiments, and the best fusion location and method are found.

(3) A spatiotemporal convolution algorithm is proposed, which connects the improved dual-stream network and the R(2 + 1)D network based on ResNet in series. The spatiotemporal feature map extracted by the dual-stream network is again input into R(2 + 1)D for spatiotemporal modeling. In order to achieve series connection, R(2 + 1)D-34 and R(2 + 1)D-50 and are used to reorganize them. This study uses the pretrained network on the ImageNet and kinetic datasets to perform experiments and fine-tunes the weights on the datasets UCF101 and HMDB51, respectively. Compared with other classic methods, the algorithm proposed in this study has achieved higher recognition accuracy.

(4) This study improves and combines several existing algorithm frameworks based on deep learning and proposes a spatiotemporal convolutional network algorithm framework for long-term videos that are difficult to process. Although comparing some classic algorithms with UCF101 and HMDB51, there is a significant improvement in accuracy, but there is still a lot of room for improvement. For example, designing the optimal network architecture is tried for 2D Conv in video recognition learning, and other datasets are added for experimentation.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, "Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *Proceedings of the 23rd ACM international conference on Multimedia*, Brisbane Australia, 2015.

[2] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1506–1515, Las Vegas, NV, USA, 2016.

[3] L. Jin, S. Gao, Z. Li, and J. Tang, "Hand-crafted features or machine learnt features? together they improve rgb-d object recognition," in *Proceedings of the 2014 IEEE International Symposium on Multimedia*, pp. 311–319, Taichung, Taiwan, 2014.

[4] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: a new representation for rgb-d based action recognition with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, Honolulu, HI, USA, July 2017.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the International conference on computer vision*, December 2015.

[6] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: deep networks for video classification," in *Proceedings of the Conference on computer vision and pattern recognition*, June 2015.

[7] Y. Zhou, X. Sun, Z. J. Zha, and W. Zeng, "Mict: mixed 3d/2d convolutional tube for human action recognition," in *Proceedings of the Conference on computer vision and pattern recognition*, June 2018.

[8] H. Kuehne, H. Jhuang, E. Garrote, and P. T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the 2011 International Conference on Computer Vision*, November 2011.

[9] L. Bottou, "Stochastic gradient descent tricks," *Lecture Notes in Computer Science, Neural Networks: Tricks of the Trade*, Springer, Berlin, Heidelberg, pp. 421–436, 2012.

[10] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

[11] M. Shu, "The value choice of the public in the era of pan-entertainment," *New Media Research*, vol. 34-35, 2018.

[12] Z. Shi, "Beware of pan-entertainment enslaving the self," *People's Forum*, vol. 44-46, 2018.

[13] W. Jia, "Pan-entertainmentism makes entertainment a fool," *People's Forum*, vol. 50-52, 2018.

[14] Q. Jiang, Y. Zhang, S. Tan, and Y. Yang, "Recognition of students' classroom behavior based on residual network," *Modern Computer*, vol. 20, pp. 23–27, 2019.

[15] X. Jin, *Research and Implementation of Face-To-Face Classroom Intelligent Management System Based on Seetaface Face Recognition Engine*, Jiangsu University, Zhenjiang, China, 2019.

[16] A. B. Dhivya and M. Sundaresan, "Tablet identification using support vector machine based text recognition and error correction by enhanced n - grams algorithm," *IET Image Processing*, vol. 14, no. 7, pp. 1366–1372, 2020.

[17] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learing-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Transformations on Industral Infomatics*, vol. 14, no. 7, pp. 3235–3243, 2018.

[18] S. Wang, H. Chen, L. Wu, and J. Wang, "Novel smart meter data compression Method .via. staked. convolutional sparse auto-encoder," *Internate Journal of Elector Power and Energy Systems*, vol. 118, Article ID 105761, 2020.