IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 June 2024, accepted 9 July 2024, date of publication 15 July 2024, date of current version 25 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3428573

RESEARCH ARTICLE

Video Behavior Recognition Model Based on Spatial Long-Distance Modeling Combined With Time-Domain Shift

DEGANG SUN[1], **YANYU ZHOU**[1], **AND ZHENGPING HU JR.**[1,2,3]

[1]College of Information Engineering, Shandong Huayu Institute of Technology, Dezhou 253034, China
[2]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
[3]Hebei Key Laboratory of Information Transmission and Signal Processing, Yanshan University, Qinhuangdao 066004, China

Corresponding author: Degang Sun (sunhan0910@163.com)

This work was supported in part by the National Science Foundation of China under Grant 61771420, and in part by the Research Fund for Pattern Recognition Application Engineering Technology Research and Development Center of Shandong Huayu Institute of Technology under Grant 201905.

ABSTRACT During the feature extraction study, the video behavior recognition algorithm had a limited ability to extract remote target and time-motion information, resulting in unsatisfactory model classification results. To enhance the network's expression capabilities, this study proposes a video behavior recognition algorithm that combines spatial long-distance modeling with a temporal shift. To efficiently extract time-domain motion features in the 2D backbone network, the residual is coupled with the time shift module. At the same time, a narrow and long core, namely $1 \times N$ or $N \times 1$ strip pool, is introduced to make the backbone effectively capture the remote information in airspace and obtain the context relations of long-distance targets. Experiments on Something-SomethingV1 and Jester datasets achieve an average recognition accuracy of 45.82% and 96.89%, respectively. The experimental results demonstrate that the proposed algorithm can fully extract time-space features of videos, which establishes certain advantages compared with other existed behavior recognition networks.

INDEX TERMS Behavior recognition, deep learning, long distance modeling, neural network.

I. INTRODUCTION

With the advancement of artificial intelligence (AI) technology, video behavior recognition, which is widely used in security monitoring systems, smart home design, intelligent video analysis, driverless systems, and other disciplines, has progressively emerged as a vital technology in computer vision. Video behavior recognition algorithms have evolved from the initial manual feature extraction method to the current deep learning approach.

Early behavior recognition algorithms adopted a method for manually extracting features. The method begins by sampling the input video to create sampling points. Subsequently, feature extraction was conducted on these sampling points to generate manual feature descriptors. These descriptors are then encoded to produce feature vectors that are subsequently trained. The final step involves the classification of the trained feature vectors to generate the output classification result. For example, Bobick and Davis [1] proposed to use background subtraction to extract foreground contour features from videos, and constructing Motion Energy Images (MEI) and Motion History Images (MHI) to represent motion information. Yang and Tian [2] used the depth image to extract the position information of human body joint points, collected the position coordinates of human body joint points, and used the human body contour formed by it for behavior feature recognition and achieved good results. The above two methods are based on contour silhouette and joint point human contour feature extraction, both of which use the segmentation of foreground and background. However, when the background is complex, the feature extraction effect is not good. Some scholars propose extracting behavioral features by tracking human motion trajectories. Wang et al. [3] proposed a dense trajectory

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

method that uses the optical flow field of dense sampling points to obtain motion trajectory information. Then, Wang and Schmid [4] proposed an improved dense trajectory algorithm. Combining the Histogram of Orientation Gradient (HOG) and the Histogram of Optical Flow Orientation (HOF) can improve performance. Applying video optical flow to match key points helps overcome interference caused by camera changes and improves the robustness of features. Manual feature extraction methods rely on human experience and introduce complex data processing processes, which are not ideal for feature extraction in complex and changeable backgrounds, occlusions, and other challenging scenarios.

In recent years, video behavior recognition methods based on deep learning have received significant attention, and researchers have applied various neural network models to algorithms. In terms of time-space feature network models, the most classic approach is the feature extraction network based on the dual flow model proposed in the literature [5], in which the dual flow network is used to extract the spatial flow of static apparent features and the time flow of motion information, respectively. Finally, the dual flow 2DCNN network is integrated to achieve enhanced classification accuracy. Wang et al. [6] borrowed from the architecture of literature [5] and also adopted the combination of spatial flow network and temporal flow network. However, different from the previous two-stream network, the long-term video combined with the sparse sampling strategy proposes temporal segment networks (TSN) that sparsely sample multiple short clips from the video as network input. Each short clip is input into the network for preliminary prediction. Finally, the prediction and classification results of the entire video are combined, and the excellent classification accuracy of 94.20% is achieved on the UCF101 dataset. In view of the complex preprocessing of optical flow and the inability to meet real-time performance, some researchers proposed to expand the time dimension of the network to form a 3DCNN. For example, the literatures [7], [8], [9], [10], and [11] used 3DCNN to create a spatio-temporal feature extraction model. Among these, Tran et al. [7] proposed a Convolutional 3D Network (C3D) using 3D convolution, which is simple, compact, and efficient. It extracts the spatial and temporal features of video sequences using 3D convolution kernels. Huang et al. [11] designed a 2D inflated operation and a parallel 3D convolutional network architecture. 2D-INFLATED Operation was used to convert pre-trained 2D ConvNets to 3D ConvNets, avoiding the pre-training problem of video data. Converting 2D convolution operations to 3D convolution operations increases the processing of time scales. Generally speaking, the expansion of 2D convolution to 3D convolution will improve the accuracy of network classification, but the amount of computation will also be greatly increased. To this end, other dimensions beyond the time dimension are considered in the literature [12], such as frame rate, total frame length of input data, network width and depth, etc. Also, to reduce the amount of computation, Luo and Yuille [13] proposed a grouped convolution model to extract features efficiently. Since the algorithm uses optical flow to extract time-series motion information, the amount of data calculation is large and time-consuming, so it is not suitable for demanding real-time requirements. The 3D network requires a large amount of computation and high hardware requirements. In order to utilize 2D convolution to achieve the effect of 3D convolution, Zhou et al. [14] proposed multi-scale frame tuples involving long-range video frames. Through temporal inference on video frames of different lengths, spatiotemporal feature information is extracted. Finally, the fusion results are obtained, but the time and space of the algorithm are less connected. To address the challenge of more efficient spatiotemporal information extraction and fusion, Lin et al. [15] proposed the incorporation of a temporal shift module into the residual structure following the convolutional layer of the spatial feature extraction network. The frame information is fused, and the online recognition mode can be formed by only merging the channel information of the previous frame. The channel shift improves the receptive field in the time domain and enables more complex time-domain modeling. In order to softly connect the network, literature [16] proposes a gate-shift model (GSM) based on the time shift module to connect the features. The features extracted by 2DCNN are divided into two branches, which are adaptively selected by the gating unit to enter the subsequent channel shift network and spatial information extraction network. The branch fusion is used for behavior classification. Starting with the dynamics and time scale of the visual rhythm of action, Liu et al. [17] proposed the Temporal Correlation Module (TCM), which uses relevant operations to extract pixel-level fine-grained temporal dynamics for fast-paced and slow-paced actions, and considers cross-temporal dynamic interactions. Adaptive selection and enhancement of the most effective movement visual rhythm information. Using a 2D convolutional neural network combined with related modules can realize the modeling of time, which not only meets the real-time performance but also solves the problem of high hardware requirements of a 3D convolutional neural network. Of course, for video action recognition models, temporal modeling is essential, and spatial feature extraction cannot be ignored.

Scene analysis and semantic segmentation in the video are conducive to the acquisition of apparent features in video behavior recognition. It is essential to capture remote context information in the process of acquiring video features. In this regard, existing methods, such as stacking local convolution and pooling operations, Wang et al. [18] proposed a convolution module to improve the long-range modeling capability of CNNs. By computing the response of a local position as a weighted sum of features at all locations, the model achieves excellent results in video classification competitions. Alam et al. [19] used spatial pyramid pooling and decoder

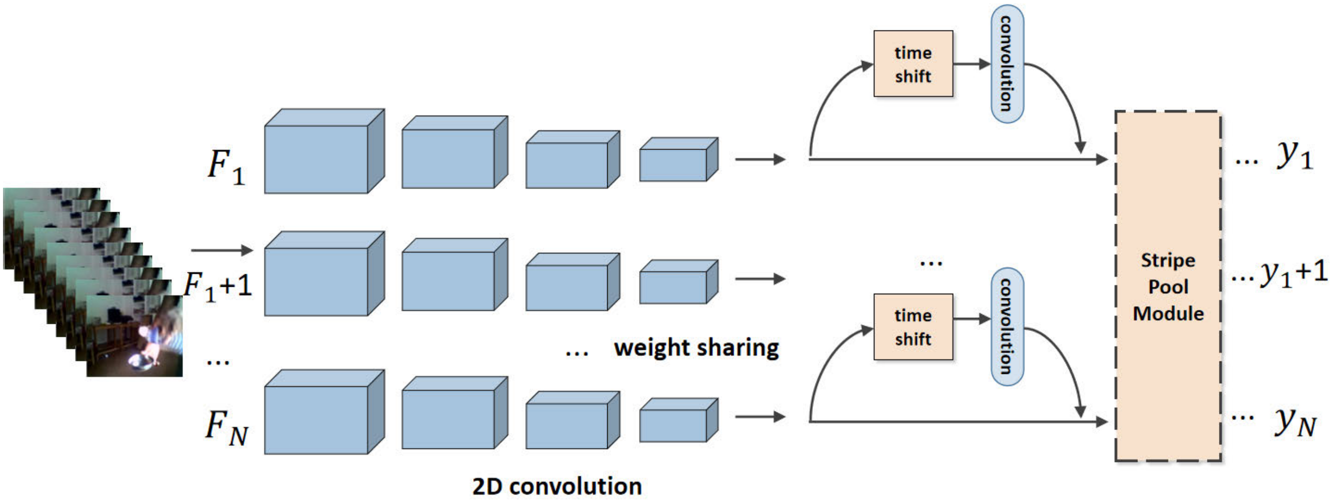


FIGURE 1. Schematic diagram of the overall model architecture.

for scene analysis. They applied depthwise separable convolution to pyramid pooling and decoder modules, resulting in a faster and more robust encoder-decoder network. He and Deng [20] proposed an adaptive pyramid pooling context network, which uses the global image to estimate the sub-region correlation coefficient and calculates the context vector with the correlation. However, these methods are limited to the input features within a square window. To capture discrete distributions that may have long-range band structures in real-world scenarios, Hou et al. [21] proposed a new band pooling strategy, which is different from traditional spatial Compared with $N * N$ pooling, it utilizes strip pools formed by long and narrow nuclei to connect long-range contexts. The fusion of spatial feature extraction modules can effectively supplement the feature extraction capability of the backbone network, enabling the backbone network to model context dependencies effectively.

To increase the network's spatiotemporal modeling capabilities and connect contextual information and long-range dependencies, this paper models video scene areas using an efficient temporal channel shift module and a stripe pooling network to extract spatiotemporal features. Due to the issue of information redundancy between neighboring video frames, this work first adopts a sparse sampling method. The visual frames are then input into the network, which expands the temporal receptive field and extracts motion information. In conjunction with the channel shift module, the channel information of neighboring frames of the video is exchanged in a specific proportion to accomplish the impact of timing information extraction. In airspace feature extraction, the strip pooling module is fused, and the narrow and long kernel $1 \times N$ or $N \times 1$ is selected to expand the airspace receptive field. Connecting the remote contexts of discrete regions in the scene, the overall network model in this paper achieves the performance of 3D convolutional networks with the complexity of 2D convolutional networks.

A. VIDEO ACTION RECOGNITION ALGORITHM MODEL BASED ON TIME-DOMAIN SHIFT AND LONG-DISTANCE SPATIAL MODELING

Figure 1 is a schematic diagram of the overall structure of the algorithm proposed in this paper. The time domain shift module is positioned before the first convolution layer of Resnet in order to augment the temporal information of the input video. The stripe pool module is situated behind a 3×3 convolution layer within Resnet's residual block. Its function is to model spatial remote information. The concurrent incorporation of the two modules into Resnet can enhance the network's capacity to model temporal and spatial remote information. The video behavior recognition process can be divided into three steps. The initial input video is subjected to preprocessing in Step 1 due to its large spatial resolution and information redundancy. This is achieved through downsampling. For a given video V, we first sample the video T frame $F_1, \ldots, F_t$, in which a certain frame is denoted as F_i. The second step involves inputting the pre-processed image frames into a two-dimensional convolutional neural network in order to extract features. This paper adopts Resnet-50 as the backbone network. Concurrently, in order to extract a comprehensive set of timing features, a portion of the extracted channel information is stored and exchanged with the adjacent frame information. The approach is implemented in the time shift module depicted in Figure 1. It is important to note that in the specific implementation, the time shift modules are connected in a residual manner to form an offline channel shift mode, which is used to extract spatiotemporal features. The strip pooling module is integrated into the feature extraction network, therby enabling further modelling of the scene region. The internal structure of the strip pool module and the backbone network also adopts the residual connection mode, whereby additional features are incorporated into the backbone extraction network without disrupting its

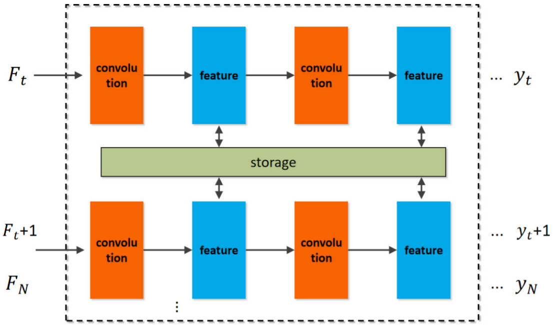


FIGURE 2. Schematic diagram of time-shift module.

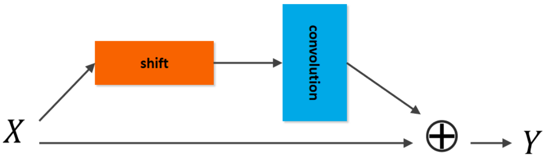


FIGURE 3. Location diagram of shift module.

functionality. The combined fusion of $1 \times N$ and $N \times 1$ bands, in the form of two channels, enhances the receptive field in a comprehensive manner. Finally, the fully-connected layer and the softmax layer must be linked to the video output, in order to obtain the video behaviour recognition and classification results.

II. TIME DOMAIN SHIFT AND SPATIAL STRIP POOLING

A. TIME DOMAIN SHIFT MODULE

In the field of video action recognition, the temporal motion information present in the video can be extracted by inputting the image sequence into a 3DCNN network. However, the memory consumption of 3DCNN is relatively high and demands advanced hardware. The parameters and computational complexity of 2D convolutional neural networks are relatively modest, but single-frame 2DCNN are unable to simulate temporal information to a greater extent. In this paper, the modeling of time is realized by exchanging the channel information of image frames.

The temporal shift module TSM is depicted in Figure 2. The data input to the network model is represented by the matrix $A \in R^{N\times T\times C\times H\times W}$, where N is the batch size, T is the number of frames, C is the number of channels, and H and W are the spatial resolutions. For the input image frame Ft, use convolution to extract features, move S channels in the feature channel number C by $+1$ and S channels by -1, and the remaining channels do not move. The choice of the shift scale hyper-parameter S is discussed in a subsequent section. The temporal shift module combines multi-frame channel information with zero parameters and zero computational cost. Although traditional 2DCNN is used, the time domain $Ft - 1$ frame, Ft frame, and $Ft + 1$ frame image information fusion is capable of effectively obtaining temporal motion information.

Assuming a conventional convolution operation with a kernel size of 3, the weight of the convolution is $w = (w_1, w_2, w_3)$. The input is an infinitely long one-dimensional vector X. The convolution operation, represented by the equation $Y = Conv(W, X)$, can also be expressed as $Y_i = w_1x_{i-1} + w_2x_i + w_3x_{i+1}$. This equation demonstrates that the convolution operation can be decomposed into two steps: a shift operation and a multiply-accumulate operation. The shift is shown in equation (1).

$$X^{-1} = X_{i-1}X^0 = X_iX^{+1} = X_{i+1}. \tag{1}$$

The superscript in equation (1) indicates a shift operation. The values -1, $+1$, and 0 indicates, respectively, a shift to the left, a shift to the right, and no shift operation. The subscript represents the result of the shift operation, X_i to the X_{i-1} position after a shift to the left, X_i to the X_{i+1} position after a shift to the right. The multiplication and accumulation operation can be expressed as:

$$Y = w_1X^{-1} + w_2X^0 + w_3X^{+1}. \tag{2}$$

The first shift operation only needs to move the address pointer when it is implemented, without multiplying and moving data. Compared to the basic 2DCNN model, the time domain shift module in this paper combines multiplication and accumulation into 2D convolutions. Although it will occupy a certain amount of memory, it will not increase the amount of extra computation. This operation can be equivalent to a temporal convolution with a convolution kernel of 3. Temporal shifting cannot move most of the channels like spatial shifting. Excessive channel number shifting not only consumes more memory resources, but also loses current frame information and affects the spatial modeling of convolutional neural networks. The results is poor classification accuracy and high resource occupancy. In order to address these issues, two distinct methodologies have been adopted. One approach is to reduce the movement of redundant data and reduce costs by selecting the optimal channel shift number S. Secondly, in order to ensure that the feature extraction capability of the model backbone network is not affected, the temporal shift module is introduced before the convolutional layer of the backbone network in the form of a residual connection. When converting a large amount of channel information, in order to avoid damaging the spatial modeling capability of the backbone network by directly inserting modules, this paper adopts the residual structure shown in Figure 3.

B. STRIP POOL MODULE

In the context of a two-dimensional input vector, which is represented as $X \in R^{H\times W}$, where H and W represent the height and width of the space, respectively, in an average pooling layer, the spatial range that needs to be pooled is $H \times W$. Following pooling, the output y is a two-dimensional vector with a height and width of $H_0 = H/h$ and $W_0 = W/w$, respectively. In the majority of cases, the average pooling

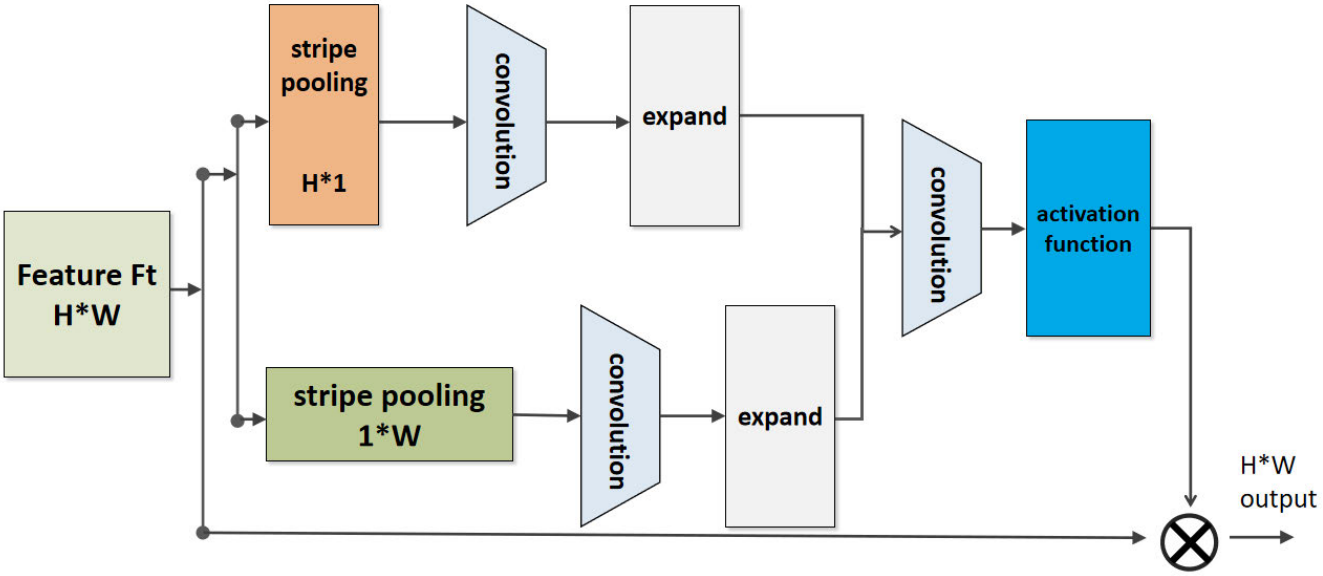


FIGURE 4. Schematic diagram of strip pool modeling speed.

operation can be expressed as:

$$y_{i_0,j_0} = \frac{1}{h \times w} \sum_{0 \leq i \leq h} \sum_{0 \leq j \leq w} x_{i_0 \times h+i, j_0 \times w+j}. \tag{3}$$

Among them, h and w are the convolution kernel sizes, $0 < i_0 < H_0$ and $0 < j_0 < W_0$, the value in the output matrix y corresponds to the value of the original matrix pooling window, and related research shows that the spatial pooling operation can successfully collect context information. However, the pooling window in general operations is a $M \times M$ square. When dealing with irregular images, irrelevant noise information is inevitably doped.

Unlike spatial pooling, for the same input $X \in R^{H \times W}$, strip pooling uses a strip pooling window whose spatial range is $H \times 1$ or $1 \times W$, and the output of strip pooling is a row or a column of eigenvalues. The output of strip pooling is the average of a row or column of feature values. The output $y \in R^{H}$ after horizontal strip pooling is as follows (4):

$$y_i^h = \frac{1}{W} \sum_{0 \leq i \leq W} x_{i,j}. \tag{4}$$

Similarly, for the output after vertical strip pooling, it is shown in the following formula.

$$y_j^w = \frac{1}{H} \sum_{0 \leq j \leq H} x_{i,j}. \tag{5}$$

Expanding the sensory field of the backbone network in the video action recognition algorithm model facilitates the understanding of the video scene. The strip pool module (SPM) composed of strip pooling is shown in Figure 4. The strip pooling module has strip pooling in both horizontal and vertical directions so that it can connect long-range contextual information from multiple directions. The input features are applied to the strip pool for context connection, and then the original path is added to achieve the purpose of long-distance modeling. In the stripe pooling module depicted in Figure 4, the input size is assumed to be $X \in R^{C \times H \times W}$, where C is the number of channels. First, feed X into two parallel passes, each containing horizontal or vertical strip pooling layers. The position and neighboring features are then adjusted using a 1D convolutional layer with a convolutional kernel size of 3, and then unfolded in their respective directions to keep the size of the original feature maps consistent. For the obtained and to get an output that contains more useful global priors, they are combined to obtain the output z as follows:

$$z = Scalse(x, \sigma(f(y))). \tag{6}$$

Among them, $Scalse()$ represents the product by site, σ is the activation function, and f represents the 1×1 convolution. In contrast to global average pooling, strip pooling establishes long and narrow ranges that can be embedded into the network to capture long-range spatial dependencies. In the event that there are ribbon structures or connections between discrete regions within the scene, the role of the band pool module is to facilitate the interactions between these discrete regions. The kernel functions for horizontal and vertical strip pooling operations included in the stripe pooling module are of considerable length and narrow width. Consequently, it is more straightforward to model the contextual relationship of distant objects within the scene area. This approach can achieve an excellent supplementary effect on the backbone network. To a certain extent, this approach avoids the capture of irrelevant area information under irregular targets by traditional spatial pooling.

In order to fully model the video scene area and to improve the temporal and spatial receptive field, the channel shift module is embedded in the form of a residual block

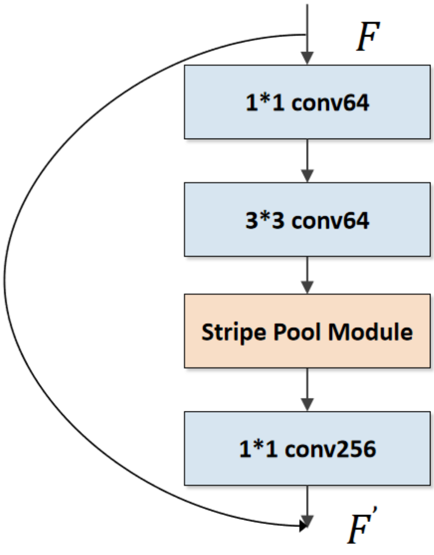


FIGURE 5. **Fusion diagram of strip pool module.**

before the convolutional layer of the backbone network. The spatial feature extraction capability of the backbone is not affected, while temporal features are still extracted. The stripe pooling module is fitted directly after the last convolution of each block of the Resnet-50 backbone, and after each 3 × 3 convolutional layers of the last block. The stripe pooling module is integrated into the ResNet-50 backbone at the 3 × 3 convolutional layer of each block. Add a stripe pooling module to a block of the Resnet-50 network as shown in Figure 5 below. The temporal shift and special stripe pooling network proposed in this paper, which combines a temporal shift module and a stripe pooling module, is referred to as the TSPNet network. The network is an end-to-end system.

III. EXPERIMENTAL SIMULATION

This research tests two action recognition datasets and compares the performance of the proposed TSPNet network. The Something-somethingV1 dataset is a large-scale labeled dataset that captures interactions between humans and items in everyday life. There are 174 kinds of behavioral acts, totaling 108,499 videos. This includes 86,017 videos for the training set, 11,522 for the verification set, and 10960 for the test set. Each video in the dataset is between 2 and 6 seconds long. The major difference from a general dataset is that the actions described by the dataset content focus specifically on time-series relationships, such as pushing something from left to right. Recognizing a video dataset is not enough to detect a target, and understanding the interactions between video contents is also crucial.

The video background in the Jester gesture recognition dataset is relatively stable, and the amount of data and categories are adequate. The dataset contains 148,092 annotated video snippets, each lasting roughly 3 seconds. The videos include 25 categories of human gestures and two categories without gestures. The video depicts a range of human movements, including swiping left or right, swiping two fingers up or down, and waving forward or back. Predicting these text annotations from video frequently requires the network to comprehend these ideas. For instance, the degree of freedom in three-dimensional space oscillates, swings, and rises.

In experiments on two video action recognition task sets, the Something-somethingV1 dataset reaches a steady state in about 50 training epochs. The initial learning rate is 0.01, the learning rate decays to 1/10 of the original every 20 epochs, and the weight decays to le-4. Using the batch stochastic gradient descent algorithm, the batch size is set to 8, and the dropout rate is 0.5. The Jester dataset is trained for about 100 epochs, in which the batch size of Jester is 6 due to limited experimental conditions, and the model is fine-tuned with weights pre-trained by Kinetics. For testing, when higher accuracy is pursued, follow the usual settings, sample 10 clips per video, and use full-resolution images at 256 for evaluation. When considering efficiency, only one video clip and an image with a resolution of 224 × 224 are used for evaluation. The experimental hardware is the mainstream NVIDIA GTX 1080TI graphics card for deep learning, and the software environment is the deep learning framework Pytorch1.1.

For complex video data, the processing method proposed in our manuscript follows these steps: First, the video is processed as a continuous image sequence. Due to the large space size of the original image, it will bring more computing costs. Therefore, we downsample the image sequence to obtain a smaller space size. The down-sampled image sequence is then fed into the Resnet network. To enhance the modeling ability of time information, a time-domain shift module is added to the convolutional layer of the Resnet network to enhance the representation of time information in the input image. After the input image sequence passes through the time-domain shift module and the convolution layer, the preliminary feature map is obtained. These feature maps will be further extracted by residuals in the Resnet network. A stripe pooling module is added behind the 3 × 3 convolution layer of residuals to enhance spatial remote information. Finally, after the characteristics of the above processing process, the video data is classified through the full connection layer and the softmax layer to complete behavior recognition.

A. SOMETHING-SOMETHINGV1 DATASET

In this section, we first experimentally verify the classification accuracy of the proposed TSPNet network on the Something-somethingV1 dataset. The TSPNet network uses Resnet-50 as the backbone to fuse the shift module and the strip pool module. Table 1 shows the experimental settings and average precision comparison between this algorithm and various mainstream algorithms on the Something-somethingV1 dataset.

The various advanced algorithms in Table 1 are consistent with the algorithm proposed in this paper, and all use

TABLE 1. Comparison table of algorithms in Something-somethingV1 dataset.

Model	Backbone Network	Pre-Training	Accuracy(%)
TSN[6]	BNInception	ImageNet	19.5
TSN[6]	ResNet-50	ImageNet	19.7
TRN-Multiscale[9]	BNInception	ImageNet	33.6
ECO[22]	BNInc+3D Res18	Kinetics	41.4
NL I3D[18]	3D ResNet-50	Kinetics	44.4
MFnet-C50[23]	ResNet-50	Kinetics	40.3
TSPNet(our)	ResNet-50	Kinetics	45.8

TABLE 2. Experimental results of different design models of Something-somethingV1 dataset.

Model	Top1(%)	Top5(%)	Top15(%)	Top20(%)
	Accuracy	Accuracy	Accuracy	Accuracy
Resnet50	17.88	43.92	68.27	68.27
Resnet50+SPM	18.21	44.19	68.78	68.78
Resnet50+TSM	44.07	73.31	88.40	88.40
Resnet50+SPM+TSM	44.32	74.00	89.13	89.13

only RGB frames as network input. Compared with the basic network model of TSN [6], the model proposed in this paper is improved by about 25%. Compared with 2D convolutional neural network models such as Non-local and TRN, it still shows great advantages in the case of low frame number input. Compared with 3D backbone extraction networks such as ECO and I3D, the present algorithm utilizing the 2D backbone networks also has higher classification accuracy. The TSPNet algorithm model completes spatio-temporal information extraction, leading to improved behavior recognition and classification accuracy.

The first set of experiments shows the classification accuracy of the TSPNet network on the Something-somethingV1 dataset. The second set of experiments will verify the effectiveness of each module of the network, such as the channel shift module and the strip pooling module. Table 2 below is the experimental data corresponding to the CMC curve on the Something-somethingV1 dataset.

High-efficiency clipping is used in the experimental test. The second row of the algorithm in the table is the Resnet-50 and the strip pool module SPM fusion network. The third row is the Resnet-50 and the time domain shift module TSM fusion network, and the fourth row is Resnet-50 Integrate with TSM and SPM modules. According to the experimental results, it can be seen that the Top1 accuracy of the network with SPM on something-something V1 dataset is about 0.3, 0.7, and 0.8 percentage points higher than that without the SPM module. Compared with no TSM module added, Top1, Top5, and Top20 increased by about 16.1, 29.8, and 20.3 percentage points, respectively. Adding the two modules of TSM and SPM, the accuracy rates of Top1, Top5, and Top20 are up to 44.32%, 74.00%, and 89.13%. The research shows that adding the stripe pool module enhances the network's ability to extract long-range context information, thereby improving the accuracy of the final classification. Regardless of whether the channel shift module is added to exchange adjacent frame information based on Resnet-50 or Resnet-50+SPM, the timing information can be effectively extracted, which significantly improves the classification accuracy.

It can be seen from the two sets of experimental results on the Something-somethingV1 dataset: (1) For behavior recognition pre-training dataset, the large-scale image classification dataset ImageNet or the behavior dataset Kinetics are generally selected. The backbone network is chosen from deep neural networks, such as the Resnet series and Inception series, as well as their deformations, such as 3D patterns.

(2) It can be seen from the second set of experiments that the addition of the SPM module and the TSM module can improve the network's classification accuracy to a certain extent. The introduction of the SPM module can effectively model long-distance contextual relationships and complement the backbone network. The TSM module extracts timing information by exchanging adjacent frame information to expand the temporal receptive field, while the channel shift module significantly extracts features.

(3) Compared with other mainstream algorithms, the TSPNet network with spatiotemporal feature fusion has certain advantages in the classification accuracy of this dataset. Compared to the 3D CNN network method, TSPNet is based on the 2D CNN architecture, which is less costly in terms of the number of parameters and the amount of computation. In addition, TSPNet has certain advantages in recognition accuracy. The relevant content is stated in the conclusion of our manuscript.

B. JESTER DATASET

On the gesture behavior recognition dataset, this paper sets up three groups of experiments. The basic conditions of the experiment are roughly the same as the Something-somethingV1 dataset. After fine-tuning the network, the model is saved, trained, and tested using an average of 3 clips. First, the initial set of experiments verifies the classification accuracy of TSPNet on the Jester dataset. This involves utilizing a spatiotemporal long-distance modeling network that combines channel shifting and strip pooling. The study also includes a comparison with existing algorithms using this dataset.

From Table 3, it can be concluded that the TSPNet proposed in this paper can achieve a classification accuracy of 96.89% on this dataset. The accuracy rate of Multiscale TRN when choosing 10-crops during the test is still about 1.5% different from that of TSPNet. Additionally, the SlowFast frame rate network with 3DResnet-50 as the backbone network is about 2% higher. Compared with 2D multi-scale

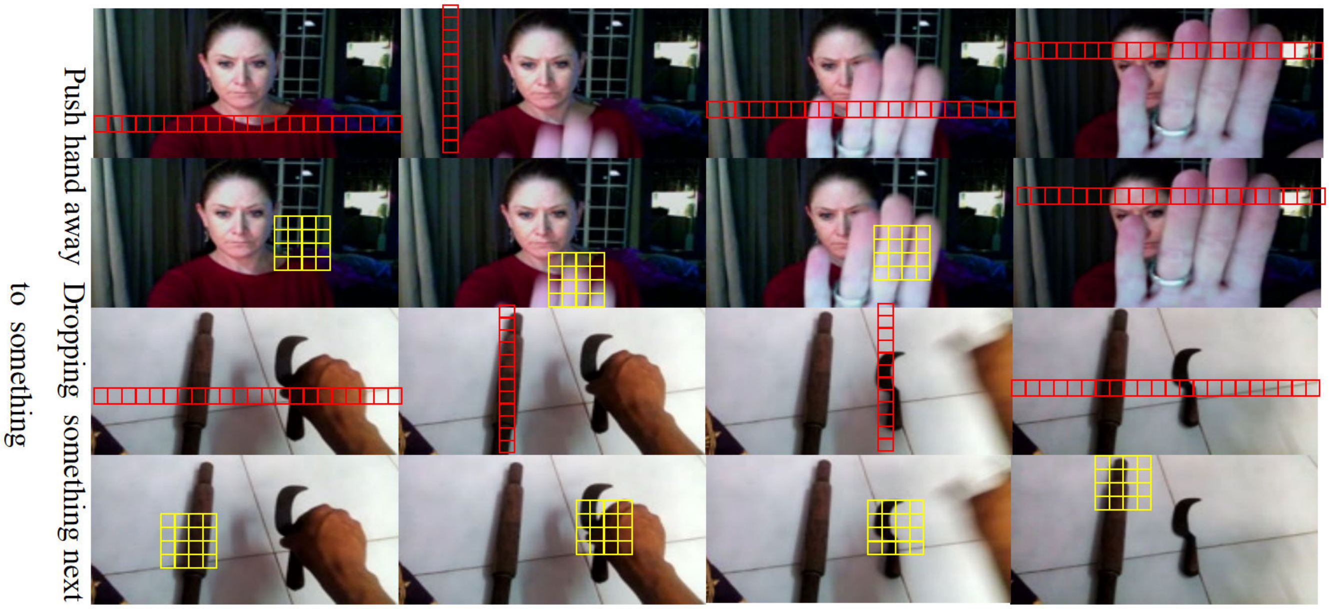


FIGURE 6. Schematic diagram of strip pool modeling speed.

TABLE 3. Performance of algorithms on the Jester dataset.

Model	Backbone Network	Top1(%)Accuracy
3D-Resnet101[24]	3DResnet-101	85.98
ECO[22]	BNInception-4a+3DRes	93.82
MFNett[23]	BN Inception	96.68
SlowFast[26]	3DResnet-50	94.46
TPR[25]	BN Inception	95.34
Multiscale TRN(10 crops)[9]	BN Inception	95.31
TSPNet(ours)	Resnet-50	96.89

TABLE 4. Experimental performance on a dataset with or without a pool of strips.

Data Set	Whether to use the stripe pool module	Parameter(M)	Top1(%)Accuracy	Top5(%)Accuracy
Jester	No	23.56	95.59	99.80
Jester	yes	32.39	95.61	99.80

time series inference models such as TRN and ECO, the module has a better effect on extracting spatio-temporal information and achieves higher classification accuracy.

The second set of experimental settings validates the effect of the strip pool module on the overall network. The feature extraction is carried out through the basic network fusion channel shift module. The experiment is carried out under the condition of keeping a single variable. The Table 4 shows the Top1 and Top5 recognition accuracy of the network with and without strip pooling on the dataset.

Compared with the mainstream algorithms, the TSPNet network has excellent accuracy. From Table 4, it can be concluded that adding the stripe pool module increases the amount of parameters to a certain extent. The 3DResnet-50 network model parameter is 48M, which is much less than the model parameter of this algorithm. But the network in Resnet-50+TSM mode already has a relatively good performance. Adding the strip pooling module on the Something-somethingV1 dataset will increase by about 0.3%. The network classification accuracy of adding the strip pooling module on the Jester dataset is the same. Based on this, the performance on the Jester dataset does not indicate that the strip pooling module does not extract long-range contextual features. This paper mainly considers the difference between the two datasets. For the Something-somethingV1 dataset, it pays more attention to the temporal relationship and interaction relationship. The extraction of long-distance context information is beneficial to the feature extraction of long-distance interaction relationships. The simple schematic diagram is shown in Figure 6. However, the background of the Jester dataset is relatively single, without obvious long discrete regions, and lacks interaction. Therefore, there are differences in the improvement of network accuracy.

The third group of experiments in this section sets the effect of different channel shift ratios on the network model. First, Resnet-50 is selected as the primary network, and a single variable is maintained for experiments. The sum of the two-way motion ratios is set to 0, 1/16, and 1/4, respectively. The classification accuracy on the Jester dataset is shown in Figure 7 below.

The channel shift ratio depicted in the figure is calculated according to the sum of the bidirectional shift ratios of adjacent frames. It can be demonstrated that the model classification accuracy rate is 82.08% when the channel shift

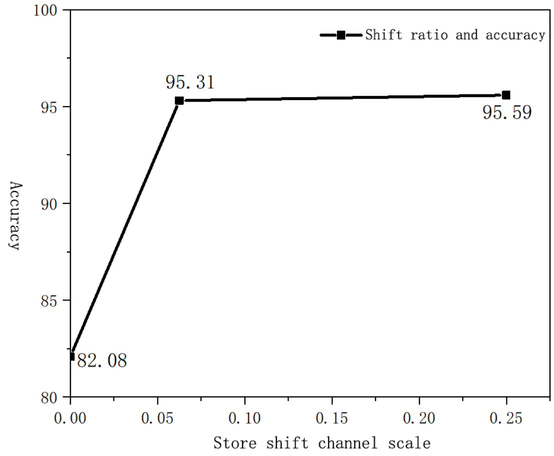


FIGURE 7. Chart of channel shift versus accuracy on the Jester dataset.

module is not used. Furthermore, when the channel shift ratio is only 1/16, the accuracy rate increases to 95.31%. As the shift ratio increases, the classification accuracy also increases to a slight degree. When the shift ratio is 1/4, the accuracy rate is 95.59%. However, it is not the case that the higher the shift ratio, the better. Furthermore, the corruption of channel information is also a consequence of out-of-memory conditions.

In this experiment, a 1/4 shift ratio is selected as the basic setting. In comparison to the shift ratio, the presence or absence of the shift module exerts a more pronounced influence on the network. The channel shift module extracts effective timing features by exchanging some adjacent frame information, thereby improving the classification accuracy.

IV. CONCLUSION

To enhance the geographical and temporal scope for the extraction of long-range context information, this research proposes a video behavior recognition TSPNet network that integrates temporal features with long-range spatial modeling. The process of channel shift is adopted to extract temporal and action-related information from video data, which is subsequently exchanged with adjacent frame picture information following the extraction of pertinent features. The narrow and long kernels of a strip-pooled variety are conducive to the extraction of contextual and long-range spatial information. The experimental results on the Something-somethingV1 and Jester datasets demonstrate that the network with the temporal shift module is capable of efficiently extracting temporal motion information, offering a significant advantage over the basic network. The SPM strip pool module performs a comparable function in the network. The integration of the SPM into the network enhances the classification accuracy of the data set that contains interactive information. The video action detection method adopts time series characteristics and long-distance spatial modeling to extract a comprehensive set of spatiotemporal data. The utilization of strip pools and channel shifts can enhance the accuracy of video action identification, providing a foundation for future research into video action recognition systems. The proposed method is constrained by its reliance on a two-dimensional convolutional neural network (CNN) architecture, which does not fully leverage the latest research methodologies, such as transformer structures. Moreover, the module under investigation in this research displays a residual structure and lacks any structural innovation. Future research should address the following areas for potential enhancement: to begin, it is possible to examine the advances that have been made in the field of residual module structure. One such advance is the grouping of convolution and multi-scale convolution. Secondly, the Transformer structure can be adopted to integrate global data at the network's edge. It is important to note that the Transformer structure will result in increased computing costs. Furthermore, it would be beneficial to investigate the potential of incorporating 3D CNN modules into the 2D CNN design, with the objective of enhancing the capacity to extract space-time data.

REFERENCES

[1] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[2] X. Yang and Y. Tian, "Effective 3D action recognition using EigenJoints," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 2–11, Jan. 2014.

[3] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.

[4] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[5] S. Karen and Z. Andrew., "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[9] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[10] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5534–5542.

[11] Y. Huang, Y. Guo, and C. Gao, "Efficient parallel inflated 3D convolution architecture for action recognition," *IEEE Access*, vol. 8, pp. 45753–45765, 2020.

[12] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 200–210.

[13] C. Luo and A. Yuille, "Grouped spatial–temporal aggregation for efficient action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5511–5520.

[14] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 831–846.

[15] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7082–7092.

[16] S. Sudhakaran, S. Escalera, and O. Lanz, "Gate-shift networks for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1099–1108.

[17] Y. Liu, J. Yuan, and Z. Tu, "Motion-driven visual tempo learning for video-based action recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 4104–4116, 2022.

[18] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[19] M. Alam, J.-F. Wang, C. Guangpei, L. Yunrong, and Y. Chen, "Convolutional neural network for the semantic segmentation of remote sensing images," *Mobile Netw. Appl.*, vol. 26, no. 1, pp. 200–215, Feb. 2021.

[20] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7511–7520.

[21] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4002–4011.

[22] Z. Mohammadreza, S. Kamaljeet, and B. Thomas., "ECO: Efficient convolutional network for online video understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 713–730.

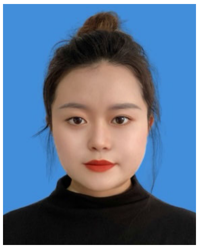
[23] M. Lee, S. Lee, S. Son, G. Pack, and N. Kwak, "Motion feature network: fixed motion filter for action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 387–403.

[24] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2874–2882.

[25] K. Yang, R. Li, P. Qiao, Q. Wang, D. Li, and Y. Dou, "Temporal pyramid relation network for video-based gesture recognition," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3104–3108.

[26] G. Kanojia, S. Kumawat, and S. Raman, "Attentive spatio-temporal representation learning for diving classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2467–2476.

DEGANG SUN was born in Liaocheng, Shandong, China. He received the B.S. degree in computer science and technology from Huanggang Normal University, Hubei, China, in 2004. He is currently pursuing the master's degree with Northeast Forestry University, Harbin, China. He is an Associate Professor with Shandong Huayu University of Technology, Dezhou. His research interest includes artificial intelligence algorithms.



YANYU ZHOU was born in Jining, Shandong, China. She received the master's degree in computer technology from Dalian Minzu University, Dalian, China, in 2023. Currently, she is a Teacher with Shandong Huayu University of Technology, Dezhou. Her research interest includes pedestrian detection.

ZHENGPING HU JR. received the Ph.D. degree in information and communication engineering from the School of Aerospace Science and Engineering, Harbin Institute of Technology, Heilongjiang, China, in 2007. Currently, he is a Professor of electronic and communication engineering with the School of Information Science and Engineering, Yanshan University, Qinhuangdao, China. His research interest includes modern digital image processing.