Hindawi
Journal of Sensors
Volume 2021, Article ID 8864870, 10 pages
https://doi.org/10.1155/2021/8864870

Research Article

RGB-D Human Action Recognition of Deep Feature Enhancement and Fusion Using Two-Stream ConvNet

Yun Liu, Ruidi Ma, Hui Li, Chuanxu Wang, and Ye Tao

College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266000, China

Correspondence should be addressed to Hui Li; lihui@qust.edu.cn

Received 17 July 2020; Revised 2 December 2020; Accepted 24 December 2020; Published 7 January 2021

Academic Editor: Giovanni Diraco

Action recognition is an important research direction of computer vision, whose performance based on video images is easily affected by factors such as background and light, while deep video images can better reduce interference and improve recognition accuracy. Therefore, this paper makes full use of video and deep skeleton data and proposes an RGB-D action recognition based two-stream network (SV-GCN), which can be described as a two-stream architecture that works with two different data. Proposed Nonlocal-stgcn (S-Stream) based on skeleton data, by adding nonlocal to obtain dependency relationship between a wider range of joints, to provide more rich skeleton point features for the model, proposed a video based Dilated-slowfastnet (V-Stream), which replaces traditional random sampling layer with dilated convolutional layers, which can make better use of depth the feature; finally, two stream information is fused to realize action recognition. The experimental results on NTU-RGB+D dataset show that proposed method significantly improves recognition accuracy and is superior to st-gcn and Slowfastnet in both CS and CV.

1. Introduction

Action recognition has a wide range of applications in various fields such as video surveillance, medical rehabilitation, virtual reality, and human-computer interaction and plays an increasingly important role in the field of computer vision [1–5]. With the development of science and technology, various mobile devices can shoot videos with higher and higher definition, and videos occupy 80% of internet traffic. The video contains a large amount of information that still pictures cannot convey, making it one of the important data sources in the field of computer vision [6, 7].

The emergence of depth sensors such as Kinect and pose estimation algorithms such as openpose [8] makes it easier to obtain skeleton data. The excellent performance of skeleton data in motion representation, antisensor noise, and calculation and storage has attracted wide attention from researchers.

However, common action recognition methods use video or skeleton data alone as input. This paper makes full use of video and skeleton data and proposes a two-stream network framework that can use both video and skeleton data. The contributions are as follows:

(1) A two-stream framework is proposed, which can use video and skeleton data at the same time, so that two algorithms can take advantage of each other's weaknesses, which significantly improves recognition performance; (2) proposed a Nonlocal-stgcn, which can obtain the dependency relationship between a wider range of joints, provide richer skeleton point features, and can better perform recognition work; (3) proposed a Dilated-slowfastnet, which can better use the depth features of video and obtain long-distance object correlation.

This paper is divided into five sections:

(1) The first section introduces the significance and background of this paper

(2) The second section introduces the common action recognition methods and points out their shortcomings

(3) In the third, a new behavior recognition method SV-GCN is proposed, and the network framework is introduced in detail

(4) In the fourth section, a detailed experimental study on the proposed method is carried out to verify the recognition performance of the network

(5) The fifth section summarizes the research content of this paper

2. Related Works

Common action recognition methods are divided into action recognition based on a single data source and action recognition methods based on multiple data sources.

Action recognition based on a single data source can be divided into video-based and skeleton based. For video-based, Wang et al. [9] proposed temporal segment networks: towards good practices for deep action recognition (TSN), which is an improvement of two-stream network; Carreira et al. [10] proposed "Quo vadis, action recognition? A new model I3D" combines 3DCNN [11] into a two-stream framework and creates a large action dataset Kinetic; the two methods above need to extract optical flow, with large amount of calculation and slow running speed. Feichtenhofer et al. [12] proposed Slowfast networks for video recognition, which do not need to extract optical flow or pretraining, greatly improving the training speed; however, since which uses random sampling to obtain video frames needed by the network, there is a problem that important video frames are ignored, which affects recognition accuracy. To solve this problem, this paper proposes a Dilated-slowfastnet, which uses dilated convolution layers instead of random sampling layer, so that the objects in the video which appear in a longer space and a longer time can also be used by our model. The capture makes the model obtain more abundant features, and the features obtained by convolution are more representative, which can adapt to various video files and improve the robustness of the algorithm.

There are three common methods of skeleton-based action recognition: Li et al. [13], Kim et al. [14], and Ke et al. [15] representing skeleton data as pseudo graph, modeling with CNN-based method; Liu et al. [16] and Morais et al. [17] represent skeleton data as a series of coordinate vectors and use RNN-based method to model; Yan et al. [18, 19] represent skeleton data as graph structure and use GCN-based method to model. The research shows that it cannot show natural dependence between the joints of human body, but skeleton data can be more consistent with the natural structure of the human body, so the third method has been widely concerned by researchers. In the method based on GCN, according to the natural connection of human body, researchers take the skeleton as the edge and the joint as the point, construct graph structure, and carry on recognition work. Yan et al. first applied GCN to action recognition and proposed a skeleton-based convolution network (st-gcn) for motion recognition, which can make full use of the natural connection between human joints for modeling, but this method uses simple GCN modeling to ignore distant joints, which may cover important motion patterns. For example, when walking, the hands and feet are closely related. Although st-gcn attempts to use hierarchical GCN to aggregate a wider range of features, the node features may be weakened during the long diffusion process [16]. In order to solve problem, this paper proposes to add nonlocal [20] in the convolution process to obtain a larger range of interjoint dependencies and provide more abundant skeleton point features.

Action recognition methods based on multiple data sources, Fan et al. [21] proposed context-aware cross-attention for skeleton-based human action recognition, proposed a cross-attention module that can extract context information directly from original RGB video, and used it in action recognition methods based on skeleton data. However, this method can only use scene context information in the original video and cannot fully use all information in the original video and does not use GCN-based method to model skeleton data. In order to solve the above problems, this paper proposes a two-stream network framework that can make full use of two types of data, video and skeleton, to further improve recognition ability. Among them, one uses Dilated-slowfastnet to process video data (V-Stream); another uses GCN-based Nonlocal-stgcn to process skeleton data (S-Stream).

To verify the superiority of proposed RGB-D action recognition based two-stream network, a large number of experiments were performed on NTU-RGB+D dataset. The experimental results show that our method achieves advanced performance.

3. RGB-D Action Recognition Based Two-Stream Network

Combining skeleton data can solve the problem of spatial complexity and the stability of video algorithms. This paper proposes to use a two-stream framework to model two types of information to enhance recognition ability. The model includes a video-based action recognition method Dilated-slowfastnet (V-Stream, to process video), which is composed of data sample layer, slow path, fast path, and side connections, and a skeleton data-based action recognition method Nonlocal-stgcn (S-Stream, to process skeleton data), it consists of 2 nonlocal blocks and 9 st-gcn blocks.

The overall framework of SV-GCN is shown in Figure 1. For a given action sample, first extract the skeleton data; then, input video and skeleton data into V-Stream and S-Stream; finally, add the softmax scores of two streams to get fused score and predict the action label.

3.1. Skeleton Data Extraction Based on Kinect. In 2010, Microsoft launched Kinect, the input device of Xbox game console, to realize real-time interaction between games and users. Computer vision researchers have found that Kinect can provide RGB-D information of the captured content and can directly provide three-dimensional bone point information, and the cost is low, making Kinect camera widely used in the field of computer vision. Kinect camera is composed of cameras, microphone, and depth sensor; cameras can emit special infrared rays, which makes image

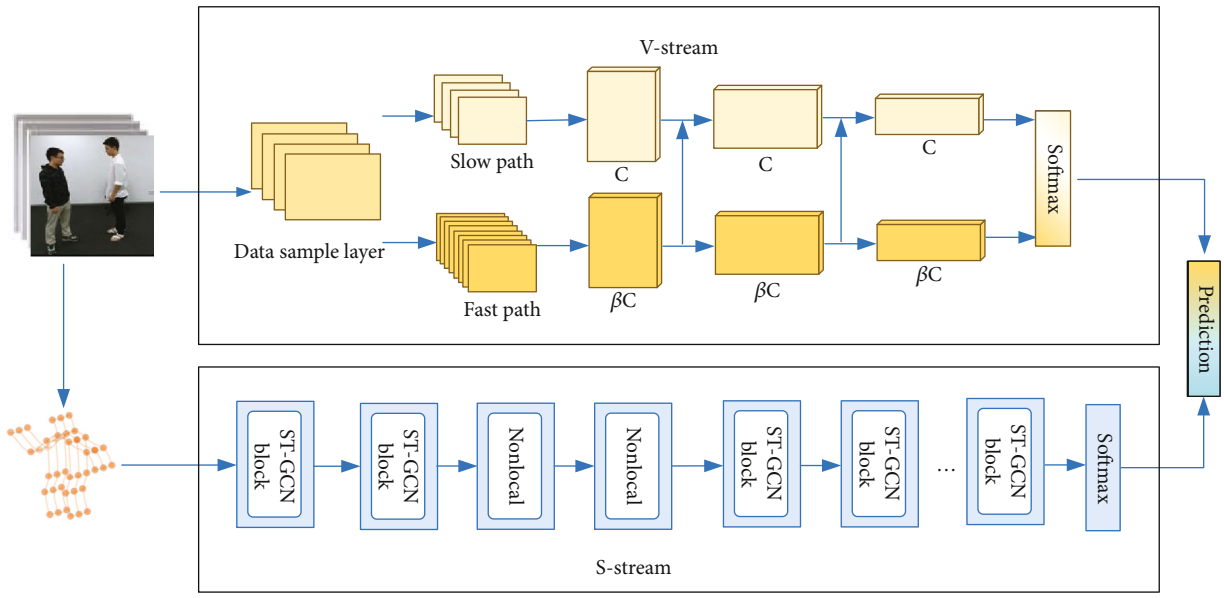

Figure 1: The overall framework of SV-GCN.

information taken by Kinect become a depth file. Each pixel uses different colors to represent the distance between the object and the camera, and the closer to the camera, the brighter the color is. Kinect can provide RGB image, 3D bone point information, depth image information, and audio signal at the same time.

There are more than 200 bones in the human body. If all the bones are modeled, a complex model will be obtained, and the calculation amount of the subsequent algorithm will be increased. In this paper, we can use the simplified joint model to extract the joint coordinates of 20 and 25 points by using Kinect (see Figure 2).

Kinect uses light coding technology, which uses laser speckle to encode a three-dimensional volume code for the entire shooting environment. Laser flash spot has strong randomness, and different patterns are produced according to different shooting distance. Therefore, using this technology, Kinect will first locate the light source in the whole space, and then when an object enters the shooting environment, it will generate a unique flash pattern and then obtain the three-dimensional position of the object according to the pattern.

The Kinect camera can detect up to six people at the same time, but only can provide two complete images of bone points. The process for Kinect camera to obtain bone points is as follows: (1) Kinect camera emits a special infrared ray to locate the whole shooting environment, calculates the phase difference according to the reflected signal, and obtains the depth image of video; (2) processes the depth image with image segmentation algorithm to obtain the human foreground; (3) uses machine learning algorithm to recognize the body in the foreground image of human body and generates bone data according to the defined joint point position.

3.2. S-Stream. The st-gcn receptive field is smaller and obtains the features of neighbor nodes so that it extracts the features of the closer joints, but the features of the farther joints are ignored, and these joints may have important motion patterns. In order to solve this problem, the paper proposes to add nonlocal operation of the spatiotemporal domain in the original st-gcn to obtain a larger range of inter-joint dependencies and provide richer skeleton point features.

The Nonlocal-stgcn is a stack of st-gcn blocks and nonlocal blocks. Each st-gcn block uses GCN and TCN alternately to transform time and space dimensions; nonlocal block acts on space-time domain at the same time and can be obtained a greater range of joint dependence, as shown in Figure 3. It is composed of two nonlocal blocks and nine st-gcn blocks, and the number of output channels of each block is 64, 64, 64, 64, 128, 128, 128, 256, 256, and 256. A data BN layer is added before the first st-gcn block to normalize input data, and a global average pooling layer is executed after the last st-gcn block. The final output is sent to softmax classifier to obtain prediction.

3.3. St-gcn Block. Nonlocal-stgcn uses 9 st-gcn blocks for learning local features between adjacent joints in space and local features of joint changes in time. Each block contains a spatial convolution and a temporal convolution. Two convolution operations are used alternately to extract spatiotemporal features. The last st-gcn block is followed by softmax for final prediction. Convolution operation of spatial graph is core of st-gcn block, which constructs a simple attention mechanism by setting its own weight parameters for each block; it introduces the weighted average value of adjacent features for each joint and sets input features of all joints in a frame as $X_{\text{in}} \in R^{n \times d_{\text{in}}}$, where d_{in} is input feature dimension and is output feature obtained by convolution of spatial graph; $X_{\text{out}} \in R^{n \times d_{\text{out}}}$ is output characteristic dimension. Thus, the convolution of space graph can be defined as formula (1):

$$X_{\text{out}} = \sum_{p \in P} M_{\text{st}}^{(p)} \bullet \tilde{A}^{(p)} X_{\text{in}} W_{\text{st}}^{(p)}. \quad (1)$$

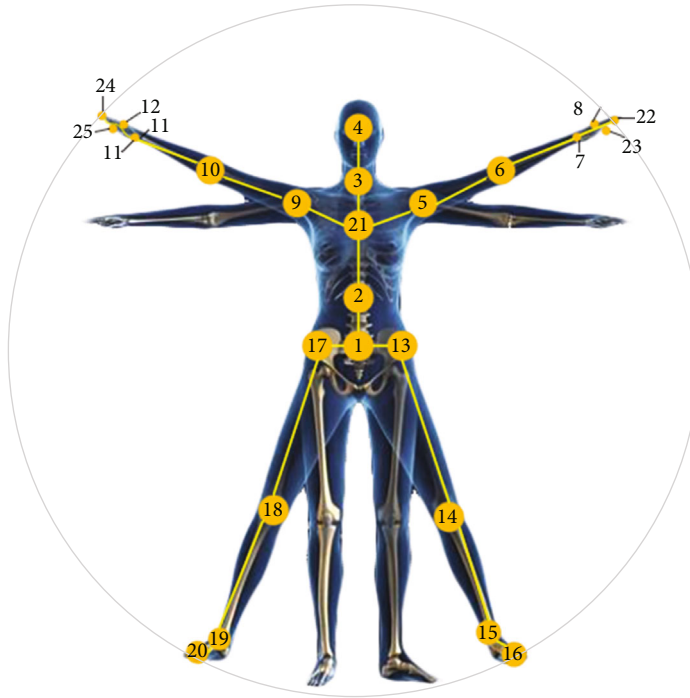

Figure 2: Schematic diagram of 25 joint points.

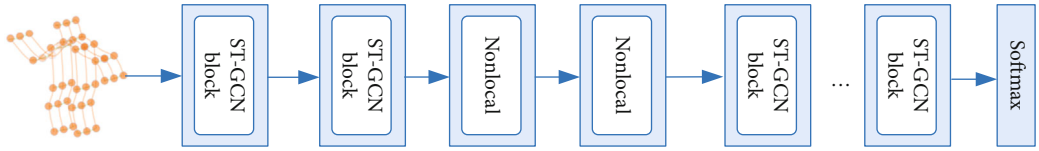

Figure 3: The architecture of S-Stream.

$\tilde{A}^{(p)} = D^{(p)^{-(1/2)}} A^{(p)} D^{(p)^{-(1/2)}} \in R^{n \times n}$ is the standardized matrix for each partition; $\bullet$ is Hadamard product; $M_{st}^{(p)} \in R^{n \times n}$和$W_{st}^{(p)} \in R^{n \times d_{out}}$ is trainable weight of each partition group, which is used to capture edge weight and feature importance, respectively.

3.3.1. Nonlocal Block. In order to solve problems of small receptive field of graph convolution network, this paper proposes to add nonlocal blocks in st-gcn. The nonlocal operation can capture correlation between long-distance pixels and realize the global receptive field of each pixel. Traditional methods usually expand the receptive field by adding convolution layer and the pooling layer, but this operation greatly increases the calculation and complexity and reduces the size of feature map. However, nonlocal can be used flexibly, placed in any position, can expand the receptive field through simple operation, and will not change the size of feature map; through different nonlocal operations, information correlation between pixels in space-time domain can be obtained.

Figure 4 shows the detailed structure of nonlocal block, $\otimes$ means the matrix multiplication, $\oplus$ is the element-wise add, and blue box indicts $1 \times 1 \times 1$ convolution. This paper uses the embedded Gaussian version, with input information x and output information y, where x and y have the same size. The implementation of nonlocal is a combination of convolution and matrix multiplication, which is defined as formula (2):

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \tag{2}$$

where $C(x)$ is the normalization parameter, in this paper, use embedded Gaussian function (formula (3)); $f(x_i, x_j)$ is a function to calculate correlation between each pixel and all position pixels. The smaller value is, the smaller influence of J position pixels on i is; it is a mapping function to calculate characteristics of point, and the smaller $f(x_i, x_j)$ value is, the smaller influence of pixels representing j position on i; $g(x_i)$ is a mapping function, which is used to calculate characteristics of point.

Embedded Gaussian is a common normalization function, a simple variant of Gaussian function. This paper considers the following forms:

$$f(x_i, x_j) = e^{\theta(x_i)^T \varphi(x_j)}, \tag{3}$$

where $\theta(x_i) = W_{\theta x_i}$ and $\varphi(x_j) = W_{\theta x_j}$ are two 1×1 convolution operations.

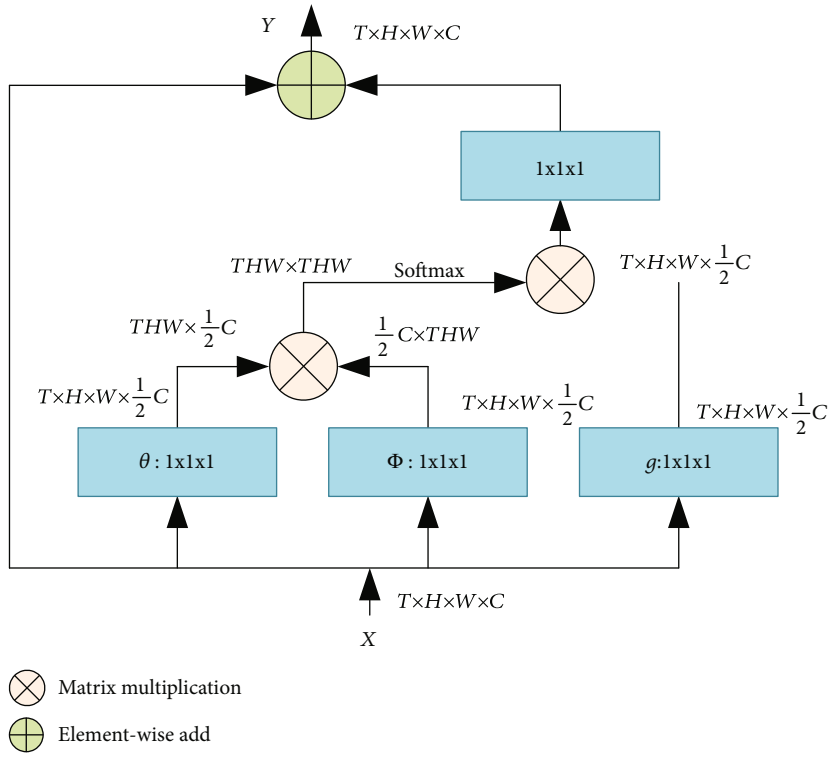

Figure 4: Nonlocal block.

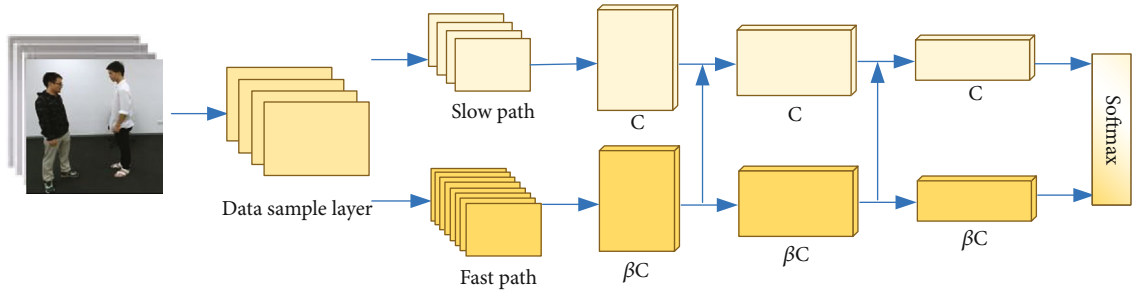

Figure 5: Dilated-slowfastnet framework.

3.4. V-Stream. Studies on the visual system of primates [22–26] have found that 80% of the visual nerve cells in primates are small cells that provide fine spatial details and colors, and 5-20% are large cells that respond to rapid time changes but are not sensitive to spatial details or colors. Slowfast uses the concept of path to reflect the analogy between small cells and large cells [12], which is composed of data sample layer, slow path, and fast path. The data sample layer uses random sampling to provide path information, which can only use the low-level features of the video and may lose video frames containing important action patterns. Dilated- slowfastnet proposes to use the dilated convolution layer to obtain video depth features in the data sample layer.

As can be seen from Figure 5, the Dilated-slowfastnet includes (I) video frame sampling layer, (II) a slow path to capture spatial semantic information, and (III) a fast path to capture fine temporal resolution motion.

Data sample layer can obtain the depth features of video, fully capture the long-distance relationship dependence, and make full use of the action mode existing in video. It is composed of batch normalization, ReLU activation function, dilated convolution, and skip connection (Figure 6), which provides different scale feature information for the two paths. For any input video clip, firstly, the dilated convolution is used to extract video features then uses normalization process to produce more stable feature distribution; in order to reduce the interdependence between parameters and alleviate the overfitting problem, the ReLU operation is performed; the introduction of convolution layer in data sample layer will increase the amount of network computation; dropout operation can eliminate some useless neurons, weaken the joint adaptability between neuron nodes, and improve the generalization ability of network; to obtain more abundant depth characteristics of video, convolution operation is performed twice and finally executes skip connection operation that makes deep and shallow features combine to produce more abundant visual features.

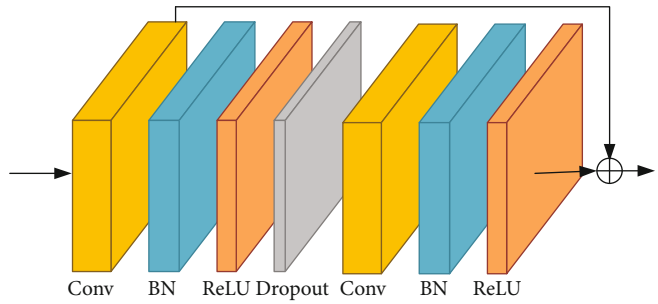

Figure 6: Data sample layer, convs represents dilated convolution, followed by BN layer and ReLU layer.

Table 1: Third-party software library version information.

Torch	NumPy	Python	Six	Tensorboardx
1.1.0	1.18.5	3.6	1.15.0	2.0
Wheel	h5py	OpenCV-python	Protobuf	Argparse
0.34.2	2.10.0	4.2.0.34	3.12.2	1.4.0
ImageIO	PyYAML	Pillow	Scikit-video	SciPy
2.8.0	5.3.1	7.1.2	1.1.11	1.4.1

Slow pathway can be any convolution model (resnet50 is used in both paths). It works in the form of spatiotemporal convolution on video clips. The key concept of slow path is to process a small amount of feature information to obtain the semantic information of video.

Fast path is a parallel path of slow path. The two paths operate on the same original segment. The feature information processed by fast pathway is time more than that of slow path (ß = 8). It is more focused on time information and is responsible for capturing the fast changing motion.

For lateral connections, the information of these two paths is integrated, so one path does not know representation learned by the other. This is achieved by horizontal connection, which has been used to fuse the two-stream network based on optical flow [27–29]. In this paper, use one-way connection to fuse the features of fast channel into slow channel. Finally, the output of each path is pooled globally. Then, two collected feature vectors are sent to softmax classifier to obtain prediction.

4. Experiment

In this section, first a large number of ablation experiments were performed on NTU-RGB+D dataset to verify contribution of added model components to recognition performance; then to evaluate performance of SV-GCN in action recognition experiments, the SV-GCN compares with previous methods.

4.1. Environment Setting. All experiments in this thesis are conducted on PyTorch deep learning framework with 6 TITANX GPUs. Among them, S-Stream uses 3 GPUs, SGD optimizer, and 100 epochs; the initial learning rate is 0.1, and the attenuation is 0.001 every 20 stages. V-Stream is randomly trained from scratch, without any pretraining; this paper reduces resolution of RGB video dataset to 480×320 pixels, uses 3 GPUs, Adam optimizer, and sets 60 epochs, and the learning rate is 0.01. For time domain, the dilated settings for the slow and fast paths are 2 and 5; for spatial domain, it randomly crops 224×224 pixels from video. The basic environment configuration of algorithm is Ubuntu16.04+ python3.6+pytorch1.1.0. Table 1 shows the version information of the third-party software library required by the algorithm.

NTU-RGB+D [30, 31] is currently the largest and most widely used action recognition dataset, contributed by Nanyang Technological University, including about 56,000 video clips of 60 types of actions. Figure 7 shows video clips of the dataset. V-Stream uses RGB video dataset; S-Stream uses 3D skeleton dataset; 3D skeleton data contains 3D coordinates of 25 human joints per frame. The dataset consists of two parts: Cross-Subject and Cross-View; performers of Cross-Subject training set and dataset are different, and perspectives of Cross-View training set and test set are different.

4.2. Training Results and Analysis. On the whole, this paper proposes a two stream model. In the training phase, V-Stream and S-Stream are trained separately, and relevant models are saved. In the test, the video and skeleton data files are input into the training model of V-Stream and S-Stream, respectively, to get the relevant test scores and then add the two to get the final prediction results. Figures 8 and 9 show loss curves during V-Stream and S-Stream training and testing. Through that, researchers find that as the number of training increases, the loss of training and testing continues to decrease. Among them, V-Stream after 60 epochs and S-Stream after 80 epochs gradually stabilize, which reflects the stability and good performance of SV-GCN.

4.3. Ablation Study

4.3.1. Two-Stream Network. The most important improvement of proposed method is simultaneous use of two data.

Figure 7: Sample videos.

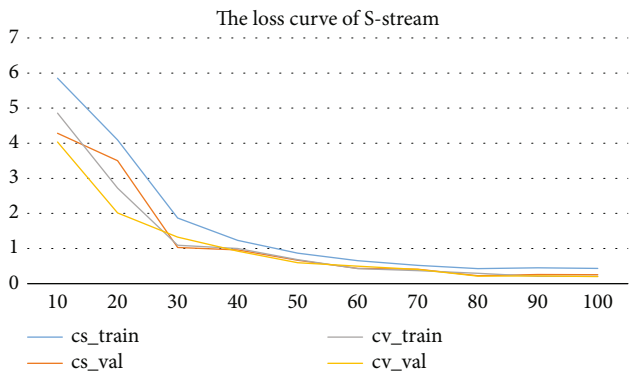

Figure 8: S-Stream loss curve of training and testing.

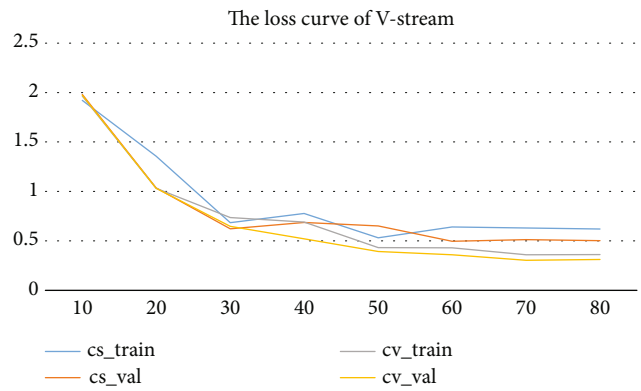

Figure 9: V-Stream loss curve of training and testing.

Table 2 shows comparison of verification accuracy of S-Stream, V-Stream, and SV-GCN. SV-GCN uses Nonlocal-stgcn and Dilated-slowfastnet to model skeleton point information and video information, respectively, fuse it by adding softmax score. The results show that SV-GCN defeated S-Stream by 1.19% and 2.24% and V-Stream by 3.39% and 1.3%. This indicates that an algorithm uses two types of data is better than one. SV-GCN can give full play to their respective advantages, make the two types of algorithms complement each other, and get better performance.

4.3.2. Nonlocal-stgcn. Since placing nonlocal blocks at different locations will result in different accuracy, in order to find the best location to place nonlocal blocks, a lot of experiments have been done in this section. The results are shown in Table 3. Among them, i-Block means to add a nonlocal block after the i-th st-gcn block. For example, 1-Block means to add a nonlocal block after the first st-gcn block. ji-Block means to add a nonlocal block to the j-th st-gcn block and i-th st-gcn block, for example, 1-2-Block means the first st-gcn block and the second st-gcn block add 1 nonlocal block. Experimental results show that adding 2 nonlocal blocks after the second st-gcn block can achieve optimal performance.

Table 4 verifies the necessity of adding nonlocal block in st-gcn to obtain a wider range of interjoint dependencies. In order to solve problems of small receptive field of GCN, this paper proposes to add nonlocal [16] to st-gcn to obtain a larger range of interjoint dependencies and provide more abundant bone point features. Experiments prove that

Table 2: Comparison of verification accuracy between two-stream network and single-stream network.

Network	CS (top 1)	CV (top 1)
S-Stream	84.32%	91.94%
V-Stream	82.12%	92.85%
SV-GCN	85.51%	94.15%

Table 3: Comparison of accuracy of adding nonlocal to different locations of st-gcn.

Network	Top 1	Top 5
Baseline	81.5%	
1-block	85.23%	98.74%
2-block	86.43%	98.62%
3-block	85.43%	97.12%
4-block	82.14%	97.2%
5-block	85.55%	97.31%
1-2-block	85.63%	96.32%
1-3-block	84.08%	95.67%
1-4-block	84.24%	92.35%
2-2-block	87.62%	97.3%
2-3-block	84.1%	95.2%
2-4-block	84.41%	94.69%
3-3-block	83.77%	94.12%
3-4-block	80.19%	91.63%
4-4-block	77.09%	91.03%
5-5-block	77.75%	90.12%

Table 4: Comparison of verification accuracy between st-gcn network and improved Nonlocal-stgcn.

Network	CS (top 1)	CV (top 1)
Nonlocal-stgcn	84.35%	91.94%
St-gcn	81.5%	88.76%

Nonlocal-stgcn after adding nonlocal module beats st-gcn by 2.85% and 3.18%.

4.3.3. Dilated-slowfastnet. The first two rows in Table 5 show the experimental results using a single path, which produce 78.05%, 88.19% and 75.32%, and 86.53% of the top 1 accuracy, respectively. The last line shows the experimental results using two paths at the same time, which is consistently better than the slow and fast only baselines.

In order to prove the necessity of using dilated convolutional layers instead of random sampling layers, this paper uses Dilated-slowfastnet and Slowfastnet to compare experiments on NTU-RGB+D dataset (Table 6). As shown, Dilated-slowfastnet beats Slowfastnet by 1.87% and 0.89%, which proves superiority of proposed method.

4.3.4. Comparison with the State-of-the-Art. To demonstrate superiority and versatility of SV-GCN, the model was compared with the latest method using NTU-RGB+D dataset.

Table 5: Slowfast fusion.

	CS (top 1)	CV (top 1)
Dilated-slow-only	78.05%	88.19%
Dilated-fast-only	75.32%	86.53%
Dilated-slowfastnet	82.12%	92.85%

Table 6: Verification accuracy comparison between Slowfastnet network and Dilated-slowfastnet.

Network	CS (top 1)	CV (top 1)
Dilated-slowfastnet	82.12%	92.85%
Slowfastnet	80.25%	93.74%

Our method divides these methods into two categories: methods based on a single data source and methods based on multiple data sources. [12, 32, 33] are action recognition methods based on RGB video; [16, 18, 19, 34–36] are based on skeleton data. Among them, results of Slowfastnet on NTU-RGB + D dataset are obtained in this environment. As shown in Table 7, the performance based on multiple data sources is better than that based on a single data source, and SV-GCN beats other methods with advantages of 1.31% and 4.85%, which proves this paper superiority of model is presented.

4.4. Analysis of Model Superiority. Video-based action recognition methods often use the traditional random sampling method to obtain key frames, but this method may lose some key action modes. In this paper, using dilated convolution layers instead of the random sampling layer can not only obtain all the action patterns but also rich context information. Graph convolution is often used as the backbone network in skeleton-based action recognition methods. However, the receptive field of graph convolution network is small; only the features of neighbor nodes can be obtained, but the remote joints are ignored, and these joints may have action patterns. In this paper, nonlocal modules are added to obtain long-distance joints dependence. And combined with skeleton data can solve the problem of spatial complexity, and video algorithm has strong stability; this paper proposes to use two stream frameworks to model the two types of information to enhance the recognition ability.

In addition, the introduction of nonlocal block, dilated convolution, and two-stream will increase the complexity and computation of the network to a certain extent. In order to improve the network performance and reduce the amount of network computing as much as possible, S-Stream does not add nonlocal after each st-gcn blocks but adds the least nonlocal blocks in the most appropriate position according to the experimental results. For V-Stream, compared with other methods, the dilated convolution can expand the receptive field without reducing the video resolution and introducing additional parameters and computation. In the two-stream information fusion stage, SV-GCN does not use the traditional multifeature fusion method but simply adds the softmax values of the two-streams, to reduce the

Table 7: Comparison of SV-GCN with other state-of-the-art methods.

Network	CS (top 1)	CV (top 1)
DSSCA-SSLM [32]	74.9%	
TCN [33]	74.3%	83.10%
GCA-LSTM [16]	76.1%	84.0%
Skelemotion [34]	76.5%	84.7%
Slowfastnet [12]	80.25%	93.74%
St-gcn [18]	81.5%	88.3%
LSTM-CNN [35]	82.9%	91.0%
Two-stream CNN [36]	83.2%	89.3%
DPRL+GCNN [19]	83.5%	89.8%
Cross-attention [21]	84.2%	89.3%
SV-GCN (ours)	85.51%	94.15%

computational complexity caused by performance improvement as much as possible.

Meanwhile, compared with the improved performance, the increased computational and network complexity are acceptable. All in all, from the view of network design structure, the network model proposed in this paper is superior to the common model, and the experiment also proves this point.

5. Conclusions

This proposes a new RGB-D action recognition-based two-stream network for action recognition task. It combines the action recognition method based on video and skeleton data so that it can not only use the semantic information provided by the former but also use the latter to solve problem of spatial complexity. At the same time, in order to give full play to the advantages of the two methods, this paper adds nonlocal in S-Stream to obtain a larger range of joint dependency and provides more abundant bone point features; in V-Stream, the dilated convolutional layers are used to replace traditional random sampling layer, which makes the algorithm better use of the depth characteristics of video to obtain representative frames, which is more suitable for action recognition task. Compared with the previous action recognition methods, this method has a great improvement. On NTU-RGB+D dataset, the model achieves the latest performance, which verifies the effectiveness of the model in behavior recognition tasks.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No. 61702295, No. 61672305) and the Shandong Province Colleges and Universities Young Talents Initiation Program (No. 2019KJN047).

References

[1] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *European Conference on Computer Vision*, pp. 103–118, Munich, Germany, 2018.

[2] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *International Joint Conference on Artificial Intelligence*, pp. 786–792, Stockholm, Switzerland, 2018.

[3] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury, "A "String of feature graphs" model for recognition of complex activities in natural videos," in *International Conference on Computer Vision*, pp. 2595–2602, Barcelona, Spain, 2011.

[4] M. R. Sudha, K. Sriraghav, S. G. Jacob, and S. Manisha, "Approaches and applications of virtual reality and gesture recognition: a review," *International Journal of Ambient Computing and Intelligence*, vol. 8, no. 4, pp. 1–18, 2017.

[5] Z. Duric, W. D. Gray, R. Heishman et al., "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1272–1289, 2002.

[6] C. Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Compressed video action recognition," in *Conference on Computer Vision and Pattern Recognition*, pp. 6026–6035, Salt Lake City, USA, 2018.

[7] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," 2018, http://arxiv.org/abs/1804.06055.

[8] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[9] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," in *European Conference on Computer Vision*, pp. 20–36, Cham, 2016.

[10] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Computer Vision and Pattern Recognition*, pp. 6299–6308, Hawaii, USA, 2017.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Conference on Computer Vision*, pp. 6202–6211, Los Angeles, USA, 2019.

[13] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *International Conference on Multimedia & Expo Workshops*, pp. 597–600, Hong Kong, China, 2017.

[14] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Computer*

Vision and Pattern Recognition Workshops, pp. 1623–1631, Honolulu, HI, USA, 2017.

[15] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Computer Vision and Pattern Recognition*, pp. 3288–3297, Hawaii, USA, 2017.

[16] Q. Li, Z. Han, and X. M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3538–3545, 2018.

[17] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Computer Vision and Pattern Recognition*, pp. 11996–12004, Los Angeles, USA, 2019.

[18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, New Orleans, USA, 2018.

[19] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Computer Vision and Pattern Recognition*, pp. 5323–5332, Salt Lake City, USA, 2018.

[20] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Computer Vision and Pattern Recognition*, pp. 7794–7803, Salt Lake City, USA, 2018.

[21] Y. Fan, S. Weng, Y. Zhang, B. Shi, and Y. Zhang, "Context-aware cross-attention for skeleton-based human action recognition," *IEEE Access*, vol. 8, pp. 15280–15290, 2020.

[22] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," *Journal of Neurophysiology*, vol. 28, no. 2, pp. 229–289, 1965.

[23] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: anatomy, physiology, and perception," *Science*, vol. 240, no. 4853, pp. 740–749, 1988.

[24] A. M. Derrington and P. Lennie, "Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque," *The Journal of Physiology*, vol. 357, no. 1, pp. 219–240, 1984.

[25] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[26] D. C. Van Essen and J. L. Gallant, "Neural mechanisms of form and motion processing in the primate visual system," *Neuron*, vol. 13, no. 1, pp. 1–10, 1994.

[27] R. P. W. Christoph and F. A. Pinz, "Spatiotemporal residual networks for video action recognition," *Advances in Neural Information Processing Systems*, pp. 3468–3476, 2016.

[28] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Computer Vision and Pattern Recognition*, pp. 1933–1941, Las Vegas,USA, 2016.

[29] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

[30] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "Ntu rgb+d: a large scale dataset for 3d human activity analysis," in *Computer Vision and Pattern Recognition*, pp. 1010–1019, Las Vegas, USA, 2016.

[31] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L. Y. Duan, and A. K. Chichung, "Ntu rgb+d 120: a large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, 2019.

[32] A. Shahroudy, T. T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in rgb+d videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1045–1058, 2018.

[33] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Computer Vision and Pattern Recognition*, pp. 156–165, Hawaii, USA, 2017.

[34] K. Papadopoulos, E. Ghorbel, D. Aouada, and B. Ottersten, "Vertex feature encoding and hierarchical temporal modeling in a spatial-temporal graph convolutional network for action recognition," 2019, http://arxiv.org/abs/1912.09745.

[35] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using LSTM and CNN," in *International Conference on Multimedia & Expo Workshops*, pp. 585–590, Hong Kong, China, 2017.

[36] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, Long Beach, USA, 2019.