Neural Computing and Applications (2023) 35:1835–1848
https://doi.org/10.1007/s00521-022-07763-8

ORIGINAL ARTICLE

3 s-STNet: three-stream spatial–temporal network with appearance and skeleton information learning for action recognition

Ming Fang[1] · Siyu Peng[1] · Yang Zhao[1] · Haibo Yuan[1] · Chih-Cheng Hung[2] · Shuhua Liu[1]

Received: 27 January 2022 / Accepted: 30 August 2022 / Published online: 5 October 2022

Abstract

Human action recognition (HAR) is one of the active research areas in computer vision. Although significant progress has been made in the field of action recognition in recent years, most research methods focus on classification of action through single type of data, and with a need to explore spatial–temporal features systematically. Therefore, this paper proposes a three-stream spatial–temporal network with appearance and skeletal information learning for action recognition, briefly coined as 3 s-STNet, which aims to fully learn action spatial–temporal features by extracting, learning and fusing different types of data. The method is divided into two consecutive stages; the first stage uses spatial–temporal graph convolutional network (ST-GCN) and two Res2Net-101 to extract the spatial–temporal features of the action from the spatial–temporal graph, RGB appearance image, and tree-structure-reference-joints image (TSRJI), respectively. The spatial–temporal graph and TSRJI image are converted from human skeleton data. The second stage fine-tunes and fuses the spatial–temporal features obtained by the independent learning of the three-stream network to make full use of the complementarity and diversity among the three output features. The action recognition method proposed in this paper is tested on the challenging NTU RGB + D 60 and NTU RGB + D 120 dataset, and the accuracy of 97.63% (cross-subject), 99.30% (cross-view) and 95.17% (cross-subject), 96.20%(cross-setup), respectively, are obtained, which achieves the state-of-the-art action recognition results in our experiments.

Keywords ST-GCN · Res2Net · Appearance · Human skeleton · TSRJI image

1 Introduction

With the rapid development of computer vision, human action recognition is widely applied to autonomous driving, intelligent monitoring, medical care, and human–computer interaction. It has become an active research topic in computer vision [1–3]. In recent years, deep neural networks have made remarkable achievements [4–6] in the field of video action recognition. Researchers began to apply deep neural networks to the field of skeleton-based action recognition. Compared with RGB video, human skeleton data contains advanced joint motion characteristics, so the representation method based on skeleton data has received great attention. Since the human skeleton structure is a natural topological map, applying convolutional neural network (CNN) [7–10], recurrent neural network (RNN) [11, 12] and their variants to skeleton action recognition cannot fully learn temporal information while requires a lot of computation. In contrast, graph

✉ Shuhua Liu
Liush129@nenu.edu.cn

Ming Fang
fangm000@nenu.edu.cn

Siyu Peng
pengsy098@nenu.edu.cn

Yang Zhao
zhaoy864@nenu.edu.cn

Haibo Yuan
yuanhb402@nenu.edu.cn

Chih-Cheng Hung
chung1@kennesaw.edu

[1] School of Information Science and Technology, Northeast Normal University, Changchun, China

[2] Center for Machine Vision and Security Research, Kennesaw State University, Marietta, GA, USA

convolutional networks (GCN) emerged as the first choice for expressing structured data. Therefore, graph convolutional network has received extensive attention from action recognition researchers. Spatial–temporal graph convolutional networks (ST-GCN) proposed by Yan et al. [1] is the first to apply graph convolutional networks to human action recognition. This method models the human skeleton and constructed it into a spatial–temporal graph with spatial and temporal information. However, this network has two drawbacks:

(1) The ST-GCN can only learn the spatial features between adjacent joints, while the network cannot learn the whole structure of human joints.
(2) The skeleton data as input leads to unsatisfactory recognition of some actions involving objects and scene information.

Hence, some researchers begin to explore on how to fully utilize the spatial–temporal information between joints [8, 9]. Caetano et al. [8] proposed a tree-structure-reference-joints image (TSRJI) which combines the skeleton tree structure and reference joint as two human skeleton processing methods. The former maintains important spatial relationships in the human skeleton by depth-first traversal of the skeleton tree, and the latter incorporates different spatial relationships between the joints. To the greatest extent in obtaining spatial information on the human skeleton, TSRJI image is introduced to complement for the lack of single spatial feature of the ST-GCN. Since both the spatial–temporal graph and the TSRJI image are skeleton data, it still cannot solve the problem of the ST-GCN which lacks the scene and object information. Compared with the skeleton data, the RGB appearance image contains rich scene information and object information, which are essential for distinguishing similar actions. Therefore, this paper proposes a three-stream spatial–temporal network combined with appearance and skeleton data for action recognition. This proposed three-stream network is briefly called 3 s-STNet. The network first converts human skeleton data into spatial–temporal graph and TSRJI image [8]. Secondly, the spatial–temporal graph, TSRJI image and RGB appearance image are taken as the input into ST-GCN, Res2Net-101, and Res2Net-101, respectively, to extract spatial–temporal features. The graph convolution achieves great success in human skeleton-based action recognition; therefore, ST-GCN is used as a basic branch of the proposed network. By inputting TSRJI image into the Res2Net-101 network to learn the spatial relationship between joints at multiple scales, this branch is abbreviated as T-R2N. The RGB image with appearance information is the input into Res2Net-101 to making full use of scene information and further improving the accuracy of recognition. This branch is abbreviated as R-R2N. Finally, the spatial–temporal features obtained by the independent learning of the three streams are fine-tuned and fused to obtain the action recognition result. Based on our theoretic reasoning and extensive experiments, the proposed model shows that the increase in scale feature representation ability can improve the robustness of action classification model. Res2Net-101 [13] has an effective multi-scale processing method and a strong ability to capture the features of static image. Therefore, this paper uses Res2Net-101 as backbone network of the other two branches. The abbreviation Res2Net-101 refers to Res2Net with a depth of 101 layers. Thus, for the simplicity, Res2Net will be used in the following discussion.

The main contributions of this paper include three aspects:

(1) The proposed 3 s-STNet for action recognition combines Res2Net and GCN, aiming to fully learn the spatial–temporal features of actions by extracting, learning and fusing different types of input data, such as spatial–temporal graph, TSRJI image, and RGB appearance image.
(2) The branch T-R2N obtains the multi-angle and multi-scale spatial features by explicitly modeling the spatial information between joints to make up for the shortcomings of the single spatial feature extraction of the ST-GCN. The other branch R-R2N stream can endow more scene and object information. In addition, all Res2Net building blocks share floating-point operations (FLOPs), such that with the increase in the receptive field of each layer while reducing the computational complexity as much as possible.
(3) The proposed model is evaluated on NTU RGB + D 60 and NTU RGB + D 120 datasets, and the recognition accuracy is improved on both datasets greatly, especially, on NTU RGB + D 120 dataset, the recognition accuracy increased by 4.47 and 3.7% with cross-subject and cross-setup standard, respectively.

The rest of this paper is structured as follows: Section 2 briefly reviews the related work and the latest progress on action recognition; Sect. 3 explains the detail of the 3 s-STNet for action recognition; Sects. 4 and 5 give experimental settings and results; the conclusions and future work then follow.

2 Related work

With the rapid development of deep learning, many action recognition models have employed deep learning methods to extract features [4, 5, 7–9, 11, 12]. Since this study

adopts both skeleton data and video sequences for action recognition, the following is a discussion of deep learning methods based on two types of data, video and skeleton, as input.

Video-based action classification. In recent years, most of video-based action recognition methods [14, 17, 18, 20] use convolutional neural network (CNN) as their baseline network. Karpathy et al. [14] used CNN to avoid the previous action confusion problems. Gan et al. [15] proposed a flexible deep CNN infrastructure, namely Deep Event Network (DevNet), that simultaneously detects pre-defined events and provides key spatial–temporal evidences. It fully demonstrates that CNN has strong robustness to RGB image. However, human action is a time-continuous image sequence, and 2D CNN does not fully learn the information between frames. Therefore, applying 2D CNN directly to the task of video action recognition will lose a lot of temporal context information. To capture temporal cues between feature maps, Lin et al. [16] proposed a temporal shift module (TSM) for processing temporal information for 2D CNN-based frameworks. However, TSM lacks explicit temporal modeling for actions such as differences among neighboring frames.

Since video action sequences are 3D spatial–temporal signals, modeling spatial–temporal information using 3D convolution is a natural and effective way. Ji et al. [17] proposed to use a 3D CNN network for action recognition. The network generates multi-channel information through adjacent input frames and performs convolution on each channel. Finally, the action features are obtained by fusing the multi-channel information. However, this model improves the recognition performance at the expense of GPU memory. With a certain amount of memory, it is easy to limit the depth of the feature map. Lu et al. [18] extracted the high-level semantic information and temporal information of video actions by using the multi-scale trajectory pooling 3D convolution descriptor (MTC3D). Due to the limitation of computing and storage capacity, this method is easy to ignore the video segmentation and feature encoding layer.

Inspired by expanding from the 2D space into 3D space–time domain, the X3D network proposed by Feichtenhofer et al. [19] was capable of incrementally expanding a tiny 2D image classification architectures along multiple network axes in space, time, width, and depth. Unlike 3D CNN, another idea for action recognition is to capture the spatial–temporal features of actions through a two-stream network. In order to better capture the motion information between still frames and continuous frames, Karen et al. [20] proposed a two-stream network that fuses spatial–temporal information. The network separately learns static RGB image and optical flow image and fused the extracted spatial–temporal features. Feichtenhofer et al. [21] proposed a two-stream architecture called SlowFast for action recognition at two different frame rates. Among them, a slow pathway operated at low frame rate to capture spatial semantics, and a fast pathway operated at high frame rate to capture motion at fine temporal resolution. The two-stream network not only demonstrates its powerful recognition performance, but also provides a good design idea for the parallel architecture of action recognition.

Skeleton-based action classification. With the successful application of CNN in the field of video action recognition, many researchers apply CNN to skeleton-based human action recognition. Cheron G. et al. [22] used appearance information and motion information to track human joints in each frame and proposed a pose-based convolutional neural network (P-CNN) descriptor. Since action recognition based on skeleton sequence is a strongly time-dependent task, a large amount of temporal information will be lost if the 2D convolutional neural network is directly applied to this task. In order to solve the above problems, recurrent neural network (RNN) is employed for processing the sequence data. Hong et al. [23] proposed a two-stream RNN model to learn the temporal and spatial features of the human skeleton. Liu et al. [24] proposed the ST-LSTM network to analyze the three-dimensional position of the skeleton joints in each frame and each processing step. Although the characteristic of RNN determines that it is suitable for dealing with temporal sequence problems, it lacks spatial modeling capabilities. Therefore, how to obtain effective spatial–temporal features has become a hot topic in this field.

In order to learn spatial–temporal information at the same time, Xie et al. [25] proposed to construct a spatial–temporal model by combining the attention RNN network and CNN network. The model firstly recalibrates the temporal information in the skeleton sequence by the attention network, and then sends the temporal information to the convolutional neural network to model the temporal and spatial information of the skeleton sequence. But the convolutional neural network of model represents the skeleton sequence as an image, that is, simply encodes the temporal information and joint information into rows and columns, and the potential-related information between the joints will be ignored. In fact, the human 3D skeleton is a natural topological map, and many studies demonstrate that graph convolutional networks (GCN) can effectively represent structured data, so researchers began to apply GCN to skeleton action recognition [26, 27].

Yan et al. [1] proposed the spatial–temporal graph convolutional network (ST-GCN), which applied graph convolutional to skeletal action recognition for the first time. In order to explore the co-occurrence relationship between the temporal and spatial domains, Chenyang et al.

[26] proposed an attention-enhanced graph convolutional LSTM network (AGC-LSTM). However, currently, only 25 joints are marked in the human skeleton data. For example, the fingers only focus on the fingertips and thumbs, which make it impossible to accurately identify similar human actions. Lei et al. [27] proposed a two-stream adaptive graph convolutional network (2 s-AGCN). First, the first-order information such as the three-dimensional coordinates of human joints and second-order information such as the length and direction of human bones are fused, and then the fused information are fed into the 2 s-AGCN model. However, the fine-grained spatial features extracted by this method are relatively low.

It can be seen from the above review that the video-based action recognition methods mainly focus on optimizing models such as CNN-LSTM, C3D, and two-stream. How to effectively perform temporal modeling has become a research problem in the field of video action recognition. Skeleton-based action recognition methods mainly focus on the modeling of human skeleton information and the optimization of graph convolution. In contrast, this paper designs a three-stream network based on the spatial–temporal graph convolutional network (ST-GCN). The network learns human skeleton information and action appearance information at the same time to make full use of the complementary and diversity between different types of data, to better explore the spatial–temporal features to improve the accuracy of action recognition.

3 Methodology

This paper proposes a three-stream spatial–temporal network (3 s-STNet) with appearance and skeleton information learning for Action Recognition, as shown in Fig. 1. The model consists of three branches, ST-GCN and two Res2Net branches, which take spatial–temporal graph, TSRJI image, and RGB video as input respectively. Among them, spatial–temporal graph [1] and TSRJI image [9] generate from skeleton sequence data. The network structure of the two Res2Net branches is the same, and both adopt a 101-layer structure. In order to distinguish between them, they are named T-R2N and R-R2N according to the input. This method first learns spatial–temporal features separately through three branches, and then fine-tunes and fuses the features to obtain the final recognition result. The three branches ST-GCN, T-R2N, and R-R2N will be introduced below in detail.

3.1 Spatial–temporal GCN (ST-GCN) stream

Yan et al. [1] applied graph convolution to action recognition first time and achieved great success; therefore, this study adopts ST-GCN as a basic branch of the 3 s-STNet network. As shown in Fig. 2, the network consists of a series of ST-GCN blocks, each block contains a spatial graph convolution and a temporal convolution. After passing several graph convolution layers and global pooling layers, the SoftMax layer classifies the actions and generates the final prediction. Therefore, for a skeleton sequence of T frames with N joints, a spatial–temporal undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ is first constructed, where the vertex set V represents all joints in the entire time, that is, $\mathbf{V} = \{\mathbf{v_{ti}} | \mathbf{t} = 1, \ldots, \mathbf{T}, \mathbf{i} = 1, \ldots \mathbf{N}\}$, the edge set E consists of the edges connecting the adjacent joints of the human skeleton in each frame $\mathbf{E_s} = \{\mathbf{v_{ti}}, \mathbf{v_{tj}} | (\mathbf{i}, \mathbf{j}) \in \mathbf{H}\}$, and connection of adjacent frames in the same joint side $\mathbf{E_F} = \{\mathbf{v_{ti}}, \mathbf{v_{(t+1)i}}\}$. Then ST-GCN applies graph convolution [28] to the constructed spatial–temporal undirected graph. Taking a single frame as an example, the connection of the joints in the skeleton is represented by the connection matrix, A, between the joints and the self-connection matrix, $\mathbf{I}$, of the joints. The formula is as follows:

$$Y_{\text{out}} = \Lambda^{-\frac{1}{2}}(A + I)\Lambda^{-\frac{1}{2}}Y_{in}W \tag{1}$$

Among them, Y_{in} is the input feature, Y_{out} is the output feature, and the diagonal matrix Λ is defined as:

$$\Lambda^{ii} = \sum_{j}\left(A^{ij} + I^{ij}\right) \tag{2}$$

The weight matrix W is the superposition of the weight vectors of multiple output channels.

On the one hand, ST-GCN obtains the spatial information of the action through the connection between the joints of human skeleton. On the other hand, it obtains the temporal information through the connection of the same joints of different frames. However, this network has two drawbacks: it can only learn the spatial features between adjacent joints, and the skeleton data as the input of the network leads to unsatisfactory recognition of some actions involving object and scene information.

3.2 T-R2N stream

In order to fully explore the spatial correlation in the human skeleton structure, this study designs the T-R2N branch. The T-R2N uses tree-structure-reference-joints image (TSRJI) as input, and Res2Net-101 as the backbone network. The extracted features are fused with those extracted by ST-GCN, which further improves the spatial learning ability of the skeleton structure. The TSRJI image and Res2Net-101 network are discussed in Sects. 3.2.1 and 3.2.2, respectively.

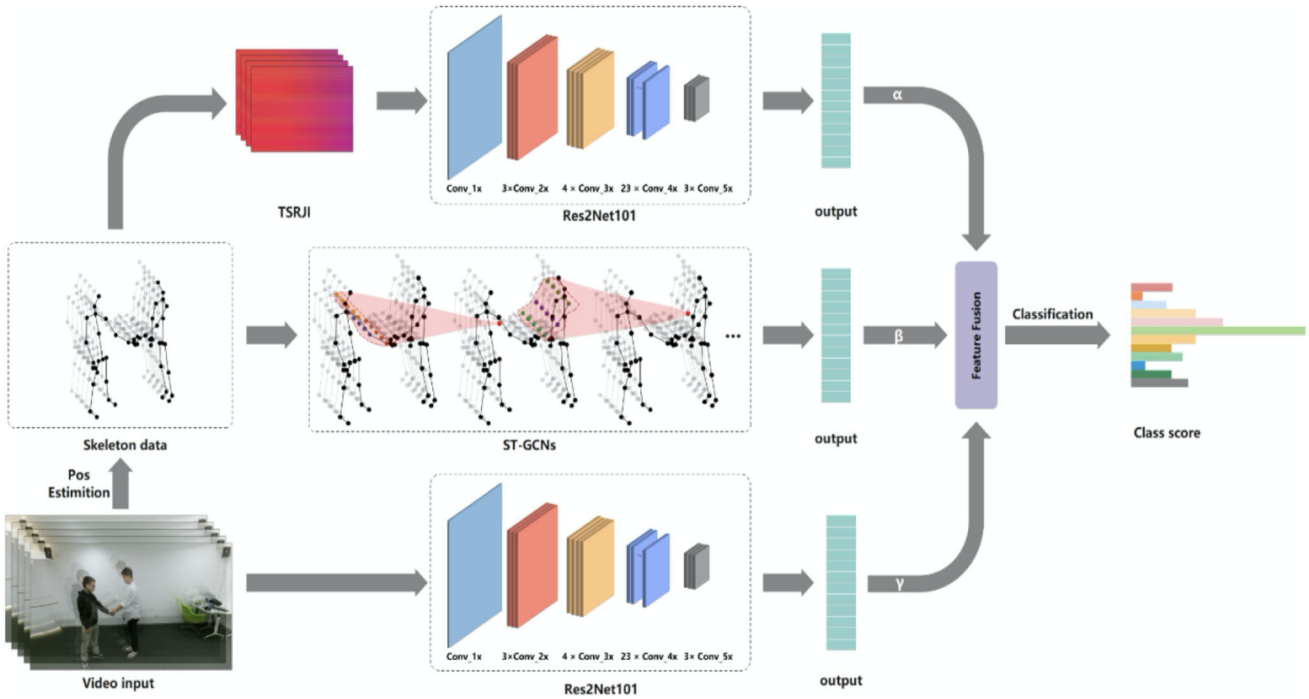

Fig. 1 Proposed 3 s-STNet framework. The spatial–temporal graph is the input for ST-GCN, RGB and TSRJI image are the input for Res2Net101

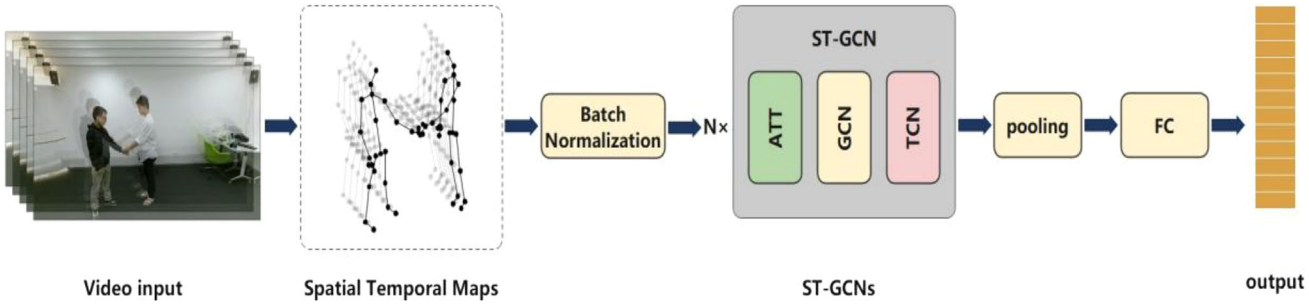

Fig. 2 ST-GCN framework

3.2.1 TSRJI image

In order to be able to effectively learn the spatial relationship between joints during human motion, Caetano et al. [9] proposed the Tree Structure and Reference Joints Image (TSRJI), which combines the skeleton tree structure [29] and reference joint [30] as two human skeleton processing methods. The former maintains important spatial relationships in the human skeleton by depth-first traversal of the skeleton tree, and the latter incorporates different spatial relationships between the joints and preserves the spatial information in the human skeleton.

Figure 3 shows the converting process from skeleton to TSRJI. Firstly, the human skeleton sequence is reordered in depth-first traversal order, and a predefined sequence S_t is generated for each frame$t$, which preserves the spatial information between the joints in the original human skeleton structure. Notation S_t regards the spatial relations between joints in the original skeleton structure as adjacent objects, effectively ensuring the minimum number of edges needed to connect a pair of joints, so that the spatial relationship between the joints maintains the tightness of the original human skeleton structure. Secondly, the literature [30] shows that the relative position between the joints can provide more useful spatial information than the absolute position, so four joints that are stable in most movements are selected as the reference joints, namely the left shoulder (9),the right shoulder (5),the left hip (17) and the right hip (13). After the depth-first traversal, the sequence S_t generates four sequences S_t^1, S_t^2, S_t^3, S_t^4, corresponding to four joints. Finally, the sequence $S_t^n (n = 1, 2, 3, 4)$ corresponding to each reference joint is connected into four

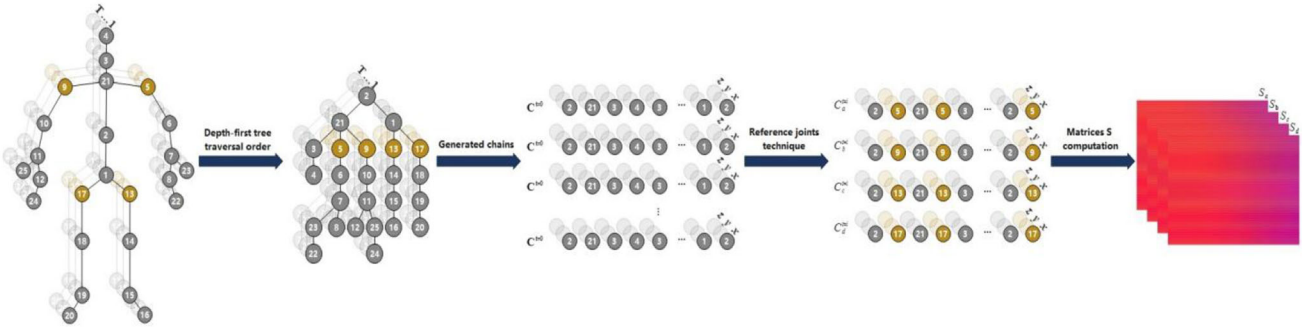

Fig. 3 A diagram shows the generation of TSRJI images

matrices M_n (n = 1, 2, 3, 4) in the order of time evolution, and the four matrices are merged to obtain the skeleton image with rich spatial–temporal information.

3.2.2 Feature learning network Res2Net

Res2Net proposed by Gao et al. [13] is a simple and effective neural network. Compared with ResNet, Res2Net replaces the single 3 × 3 convolution kernel of ResNet with residual-like connections inside the residual-block. This architecture performs multi-scale feature representation at the granularity level and increases the receptive field which is conducive to fully extracting the local and global information of the feature map.

In this study, Res2Net-101 is used to learn and classify the image features and the network structure is shown in Fig. 4. The input feature maps are divided into n subsets. Firstly, Res2Net extracts the feature information of a subset through 3 × 3 convolution, and then the convolution output and another input feature subset are input into next 3 × 3 convolution together for feature extraction. Repeat this operation until all input feature subsets are processed. In order to better fuse feature information of different scales, the processed feature map information is connected together and input to a 1 × 1 convolution for information fusion. Specifically, assuming that there is a feature map P_i (i= 1, 2,…, n), then the output Y_i is calculated as follows:

$$Y_i = \begin{cases} P_i & i = 1 \\ C_i & (P_i) i = 2 \\ C_i & (P_i + Y_{i-1}), \quad 2 < i \leq n \end{cases} \tag{3}$$

Among them, n represents the scale dimension, and C_i is a 3 × 3 convolution. When $n = 4$, the above formula can be written as:

$$\begin{cases} Y_1 = P_1 \\ Y_2 = C_2(P_2) \\ Y_3 = C_3(P_3 + Y_2) \\ Y_4 = C_4(P_4 + Y_3) \end{cases} \tag{4}$$

3.3 R-R2N stream

Although the skeleton contains advanced joint motion features, many studies show that appearance features such as objects and scenes can complement the skeleton features. Therefore, a new branch R-R2N is designed in this paper, as shown in Fig. 5. This branch takes the RGB appearance image as an input to R-R2N which is the Res2Net-101. The RGB image includes rich scene information and object information, such as books, which is essential for distinguishing between actions such as reading, writing, and typing.

4 Experimental settings

We verify the effectiveness of the proposed method on challenging action recognition dataset NTU RGB+D 60 [31] and NTU RGB+D 120 [32]. This section first briefly introduces the two datasets, their evaluation criteria, and then describes the experimental and evaluation results in detail.

4.1 Datasets

The dataset NTU RGB + D 60 and NTU RGB + D 120 are captured simultaneously by three Microsoft Kinect V2 cameras. Each sample consisted of RGB video, depth map, 3D skeleton data, and infrared video, released by Nanyang Technological University in 2016 and 2019. The NTU RGB + D 60 consists of 60 action categories performed by 40 subjects with a total of 56,880 samples. The latter is an extended version of the former, with 106 subjects collected 120 types of actions, including 114,480 video samples. In order to enrich the diversity and complexity of the data, the NTU RGB + D 120 dataset increased the number of camera views to 155. In our study, RGB video, 3D skeleton data, and TSRJI images generated from 3D skeleton data as the model input. An example is shown in Fig. 6.

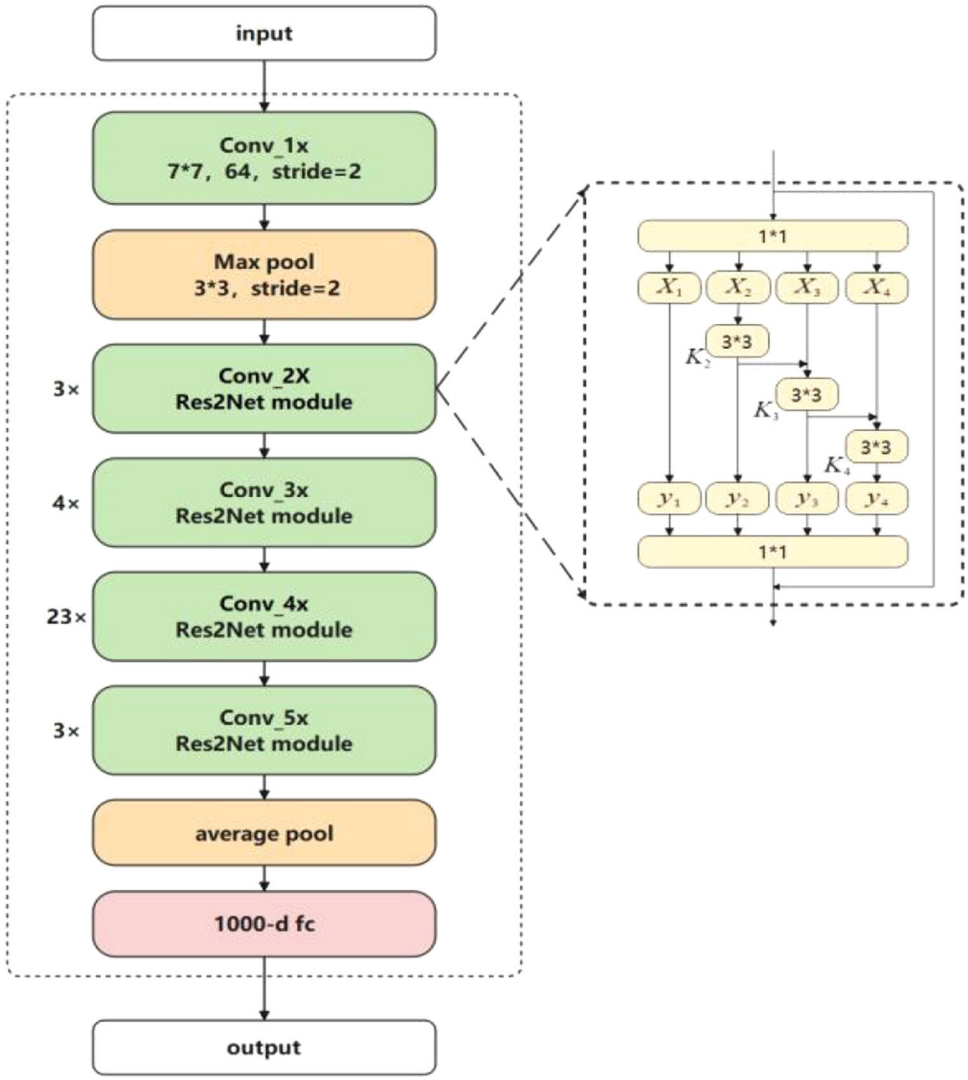

Fig. 4 Res2Net-101 architecture

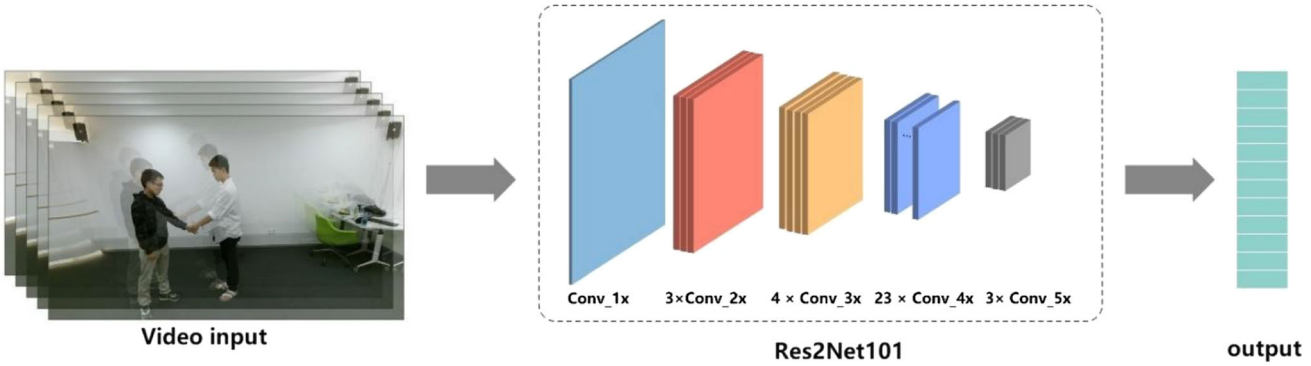

Fig. 5 R-R2N framework

4.2 Evaluation metric

NTU RGB + D 60 dataset is divided into a training set and a test set by two different standards: cross-subject and cross-view. The cross-subject standard divides 40 subjects into training and test sets by ID, each group of 20 subjects, including 40,320 and 16,560 samples. The cross-view standard is divided according to camera number, of which 37,920 samples collected by camera 2 and camera 3 are

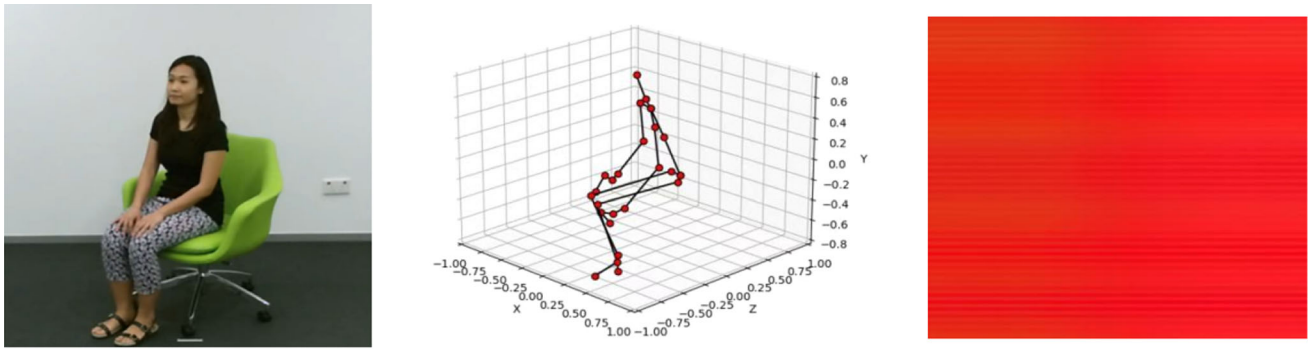

Fig. 6 Example from NTU RGB + D 60 dataset: RGB video, 3D skeleton data, and TSRJI image generated from 3D skeleton data

used as the training set, and 18,960 samples collected by camera 1 are used as the test set.

The cross-subject criterion divides 40 subjects into training and test sets by id, with 20 subjects in each group, NTU RGB + D 120 dataset is similar to the evaluation criteria used in the NTU RGB + D 60 dataset. Among them, the cross-subject standard is to divide 106 subjects into training set and test set, with 53 subjects in each group, including 63,026 and 50,922 samples. The cross-setup standard divides subjects into even and odd IDs, all samples with even identifications (IDs) are selected for training, whereas those with odd IDs are used for testing. This study strictly follows this standard for experimental evaluation.

4.3 Implementation details

The model uses the PyTorch framework, the hardware platform is Intel® CoreTM i5-8400, 64 GB memory NVIDIA 1080ti. The three branch networks of the model learn spatial–temporal features independently. Among them, the initial learning rate of ST-GCN stream is set to 0.1, the learning rate is divided by 10 every 10 epochs, and the batch size is 16. The initial value of learning rate of R-R2N stream is set to 0.1, and the learning rate is divided by 10 every 20 epochs, and T-R2N with same settings as R-R2N. CrossEntropyLoss is used as the loss function and SGD optimizer is used as the gradient descent algorithm. The RGB image and TSRJI image are input of R-R2N and T-R2N stream network. Firstly, the RGB and TSRJI image are flipped horizontally, and randomly crop them to 224×224. At the same time, the mean removal processing and color enhancement are performed for each pixel. The pixel value is transformed from $H \times W \times C$ of [0, 255] to $C \times H \times W$ of [0.0, 1.0], where H is height, W is width, and C is channel. The training set and test set use the same preprocessing method.

5 Experimental results and analysis

The experiments were first carried out on NTU RGB + D 60 dataset. After fine-tuning the proposed method, it is compared with other datasets for evaluation.

5.1 Performance of single stream

In order to achieve the best recognition effect, the model proposed is firstly fine-tuned and optimized for each branch network, and the optimized results are compared with other methods.

5.1.1 Fine-tune the ST-GCN

Under the cross-subject and cross-view evaluation indicators, the batch size is adjusted from 64 [1] to 16. The experimental results are shown in Table 1. The recognition accuracy of the fine-tuned ST-GCN network reached 83.65 and 91.59%, which are higher than the original ST-GCN network by 2.05 and 3.29%, respectively.

5.1.2 T-R2N optimization

The T-R2N network takes tree-structure-reference-joints image (TSRJI) as the input, aiming to fully learn the spatial relationship between joints by using the effectiveness of the multi-scale feature learning of Res2Net-101. Four

Table 1 Comparison of the recognition performance of ST-GCN before and after fine-tuning on the NTU RGB + D 60 dataset

Method	Accuracy (%)	
	Cross-subject	Cross-view
ST-GCN [1]	81.60	88.30
ST-GCN (ours)	83.65	91.59

reference joints of skeleton data are stable in the most of actions; therefore, this study explores the performance of different fusion methods. Table 2 shows the recognition accuracy of four reference joints fused using simple fusion, weighted linear fusion, late fusion on the NTU RGB + D 60 dataset. Among them, simple fusion stacks the skeleton image of the four reference joints, and weighted linear fusion merges the skeleton images of the four reference joints using the linear weighting. Late fusion takes each reference joint image as input to T-R2N, and then performs unweighted fusion to predict the output score. Table 2 shows the recognition accuracy of late fusion is higher than the other two fusion methods. Therefore, this paper adopts late fusion for the four reference joints.

Compared with CNN, Res2Net-101 adopts multi-scale processing method and has stronger feature extraction ability. Table 3 shows the recognition accuracy of CNN [9], ResNext-101 [33], Res2Net-50 [13] and the T-R2N proposed in this paper on the NTU RGB + D 60 dataset with TSRJI image as input. The accuracy of Res2Net-101 on cross-subject and cross-view is improved by 20.05 and 14.83%, respectively, than that of CNN [9], and achieves the best results among all networks. The experimental results show that the T-R2N network can learn the spatial relationship between joints at multiple scales and can better model the features of the skeleton image. Therefore, this paper uses Res2Net-101 to learn and classify the features of the TSRJI skeleton image.

5.1.3 Performance of R-R2N

Similar to the T-R2N branch, the R-R2N branch uses RGB appearance image as the input of Res2Net-101 to extract appearance information such as objects and scenes involved in the target person. Using the cross-subject and cross-view criteria, the recognition accuracy is 63.68 and 73.38%, respectively.

5.2 Performance of 3 s-STNet

Table 4 shows the recognition accuracy of the three single-stream networks and the fusion network 3 s-STNet on the NTU RGB + D 60 dataset, and the three branches perform linear weighted fusion. The accuracy of the 3 s-STNet is higher than the models used in our comparison. Figure 7 shows 3 s-STNet improves the accuracy of almost all action classes by fusing the three streams, which further proved the effectiveness of the method proposed.

Table 3 Recognition accuracy of CNN [9], ResNext-101, Res2Net-50 and T-R2N on NTU RGB + D 60 dataset

Method	Accuracy(%)	
	Cross-subject	Cross-view
CNN [9]	73.30	80.30
ResNext-101	93.28	94.50
Res2Net-50	93.26	94.99
Res2Net-101	93.35	95.13

Table 4 Performance comparison of single-stream network and 3 s-STNet on NTU RGB + D 60 dataset

Method	Accuracy(%)	
	Cross-subject	Cross-view
ST-GCN(Fine tuning)	83.65	91.59
R-R2N	63.68	73.38
T-R2N	93.35	95.13
3 s-STNet (without weighted fusion)	97.37	99.02
3 s-STNet (with weighted fusion)	97.63	99.30

5.3 Ablation study

The proposed 3 s-STNet is based on ST-GCN, and then adds T-R2N aiming to learn whole structure of human joints and R-R2N aiming to obtain object and scene information. Here, two ablation studies are conducted separately to demonstrate the impact of T-R2N and R-R2N on model performance.

5.3.1 R-R2N stream

The first ablation experiment is to demonstrate the importance of appearance stream R-R2N to the model, and we denote the model without the R-R2N stream as 2 s-STNet. Figure 8 analyzes the recognition performance of 2 s-

Table 2 Recognition accuracy of four reference joints fused using different fusion methods on the NTU RGB + D 60 dataset

	Input	Accuracy(%)	
		Cross-subject	Cross-view
Method	Joints-5/9/13/17 (simple fusion)	75.79	80.17
	Joints-5/9/13/17 (weighted linear fusion)	79.46	84.42
	Joints-5/9/13/17(late fusion)	93.35	95.13

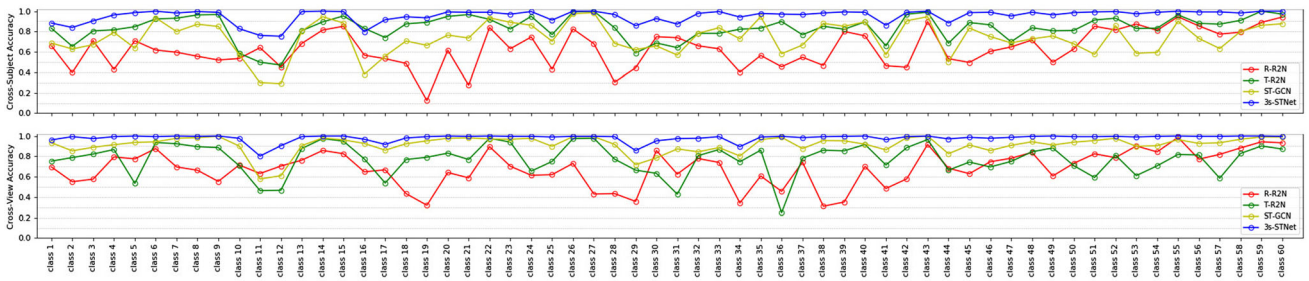

Fig. 7 Accuracy of each action class on the cross-subject index (top) and cross-view index (bottom) of the NTU RGB + D 60 dataset

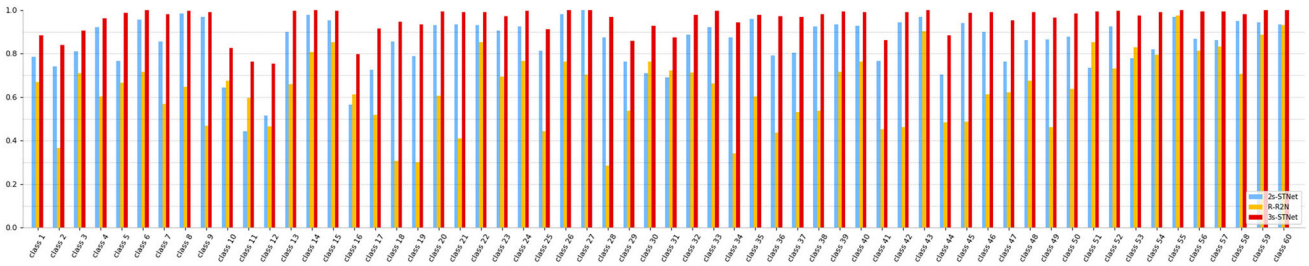

Fig. 8 Accuracy of each class of action under the cross-subject standard of the NTU RGB + D 60 dataset

STNet, R-R2N and 3 s-STNet for each class, where 2 s-STNet takes skeleton data as the input and R-R2N takes RGB appearance information as the input. Based on the data shown in Fig. 8, we can draw the following conclusions:

(1) When the action class involves object information, the recognition performance of 2 s-STNet is mostly lower than that of R-R2N, such as clapping (class 10), reading (class 11), put on a shoe (class 16), type on a keyboard (class 30), and pat on back (53). Because the skeleton data lacks the object and scene information involved in the action, the recognition effect of 2 s-STNet on these action classes is not satisfactory.

(2) 3 s-STNet improves the recognition accuracy of all action categories, and the experimental results show that the high-level joint motion features contained in the skeleton data are complementary to the low-level features (such as appearance information), and the R-R2N stream provides more scene and object information for 3 s-STNet.

Tables 5 and 6 show the performance comparison of 2 s-STNet and some advanced action recognition methods on NTU RGB + D 60 and NTU RGB + D 120 datasets. It can be seen that without R-R2N stream, our 2 s-STNet achieves the state-of-the-art recognition performance on NTU RGB + D 120; however, it is lower than VPN + + [34] on NTU RGB + D 60. The possible reason is that VPN fused different types of data, including RGB image. Therefore, our study improves the performance of the model by introducing R-R2N stream.

5.3.2 T-R2N stream

The second ablation experiment is to explore the impact of further removing T-R2N stream. As given Table 7, after the fusion of the T-R2N and the ST-GCN, the 2 s-STNet achieves higher recognition accuracy than those obtained by each component working independently. This indicates that the T-R2N obtains the spatial information between joints through multi-angle and multi-scale learning to complement for the shortcomings of the single spatial feature extraction of the ST-GCN. In addition, T-R2N stream gets higher recognition accuracy than ST-GCN. Compared with ST-GCN which only learns the spatial features of adjacent joints, the TSRJI image uses four stable reference joints to learn different spatial relations and improves the performance of network classification with explicit correlation joints.

5.4 Comparison with other advanced approaches

We compare the proposed 3 s-STNet network with advanced action recognition methods on the NTU RGB + D 60 and NTU RGB + D 120 datasets. Table 8 shows the recognition accuracy are improved by 1.03% and 0.2% than VPN + +

Table 5 Performance comparison between 2 s-STNet and the latest action recognition method on NTU RGB + D 60 dataset

Method	Year	Pose	RGB	Accuracy(%)	
				Cross-subject	Cross-view
STA-Hands [35]	2017	√	√	82.50	88.60
Altered STA-Hands [36]	2018	√	√	84.80	90.60
PEM [37]	2018	√	√	91.70	95.20
DGNN [38]	2019	√	×	89.90	96.10
Separable STA [39]	2019	√	√	92.20	94.60
P-I3D [40]	2019	√	√	93.00	95.40
NAS-GCN [2]	2020	√	×	89.40	95.70
VA-fusion [3]	2020	√	×	89.40	95.00
MS-G3D Net [41]	2020	√	×	91.50	96.20
VPN [42]	2020	√	√	95.50	98.00
HCSF [43]]	2021	√	×	91.60	96.70
VPN + + [34]	2021	√	√	96.60	99.10
2 s-STNet (ours)	–	√	×	96.48	98.74

Table 6 Comparison of 2 s-STNet and the latest action recognition method on NTU RGB + D 120 dataset

Method	Year	Pose	RGB	Accuracy(%)	
				Cross-subject	Cross-setup
Two stream Att LSTM [44]	2017	√	×	61.20	63.30
I3D* [45]	2017	×	√	77.00	80.10
PEM [37]	2018	√	×	64.60	66.90
Two-streams + ST-LSTM [32]	2019	√	√	61.20	63.10
2 s-AGCN [27]	2019	√	×	82.90	84.90
Separable STA [39]	2019	√	√	83.80	82.50
VPN [42]	2020	√	√	86.30	87.80
MS-G3D Net [41]	2020	√	×	86.90	88.40
FGCN [46]	2021	√	×	85.40	87.40
4 s-MST-GCN [47]	2021	√	√	87.50	88.80
VPN + + [34]	2021	√	√	90.70	92.50
2 s-STNet (ours)	–	√	×	93.60	94.98

[34] under cross-subject and cross-view standard on NTU RGB + D 60 dataset. Due to the robustness of the VPN ++ [34] method to view changes, 3 s-STNet outperforms it by only 0.2% under the cross-view standard. However, it is 1.03% higher than it under the Cross-Subject standard, indicating that 3 s-STNet can explore more action features involving objects. Table 9 shows the recognition accuracy increased by 4.47 and 3.7% with cross-subject and cross-setup standard on NTU RGB + D 120 dataset. It can be seen that compared to methods such as DGNN [38], NAS-GCN [2] and MS-G3D Net [41] which only use skeleton information, 3 s-STNet is able to disambiguate actions with similar visual appearance. As shown in Tables 8 and 9, the methods that use both skeleton and appearance information, such as separable STA [39], 3 s-STNet not only models the skeleton spatial–temporal features, but also improves the classification of appearance-similar actions, thereby improving the recognition accuracy. Furthermore, 3 s-STNet achieves lower computational cost compared to existing methods [1, 2, 26, 38, 48]. Although it is a little more computationally expensive than Shift-GCN [48], it improves by 6.93% for cross-subject and

Table 7 Comparison of recognition accuracy of NTU RGB + D 60 dataset before and after 2 s-STNet fusion

Method	Accuracy (%)	
	Cross-subject	Cross-view
ST-GCN (Fine tuning)	83.65	91.59
T-R2N	93.35	95.13
2 s-STNet (with weighted fusion)	96.48	98.74

Table 8 Performance comparison of the latest action recognition methods on the NTU RGB + D 60 dataset

Method	Year	Pose	RGB	FLOPs(G)	Accuracy(%)	
					Cross-subject	Cross-view
STA-Hands [35]	2017	√	√	–	82.50	88.60
ST-GCN [1]	2018	√	×	16.3	81.50	88.30
Altered STA-Hands [36]	2018	√	√	–	84.80	90.60
PEM [37]	2018	√	√	–	91.70	95.20
AGC-LSTM [26]	2019	√	×	54.4	89.20	95.00
DGNN [38]	2019	√	×	126.8	89.90	96.10
Separable STA [39]	2019	√	√	–	92.20	94.60
P-I3D [40]	2019	√	√	–	93.00	95.40
NAS-GCN [2]	2020	√	×	73.2	89.40	95.70
VA-fusion [3]	2020	√	×	–	89.40	95.00
Shift-GCN [35]	2020	√	×	10	90.70	96.50
MS-G3D Net [41]	2020	√	×	–	91.50	96.20
VPN [42]	2020	√	√	–	95.50	98.00
HCSF [43]	2021	√	×	–	91.60	96.70
VPN + + [34]	2021	√	√	–	96.60	99.10
3 s-STNet (ours)	–	√	√	31.9	97.63	99.30

Table 9 Performance comparison of the latest action recognition methods on the NTU RGB + D 120 dataset

Method	Year	Pose	RGB	FLOPs(G)	Accuracy(%)	
					Cross-subject	Cross-setup
Two stream Att LSTM [44]	2017	√	×	–	61.20	63.30
I3D* [45]	2017	×	√	–	77.00	80.10
PEM [37]	2018	√	×	–	64.60	66.90
Two-streams + ST-LSTM [32]	2019	√	√		61.20	63.10
2 s-AGCN [27]	2019	√	×	37.3	82.90	84.90
Separable STA [39]	2019	√	√	–	83.80	82.50
Shift-GCN [48]	2020	√	×	10	85.90	87.60
VPN [42]	2020	√	√	–	86.30	87.80
MS-G3D Net [41]	2020	√	×	–	86.90	88.40
FGCN [46]	2021	√	×	–	85.40	87.40
4 s-MST-GCN [47]	2021	√	√	–	87.50	88.80
KShapeNet [49]	2021	√	×	–	90.60	86.70
VPN ++ [34]	2021	√	√	–	90.70	92.50
3 s-STNet (ours)	–	√	√	31.9	95.17	96.20

2.80% for cross-view on NTU RGB + D 60 dataset and 9.27% for cross-subject and 8.60% for cross-setup NTU RGB + D 120 dataset. In summary, the proposed method achieves the state-of-the-art recognition results and with relatively low computational complexity at present.

6 Conclusions

This paper proposes a three-stream spatial–temporal network with appearance and skeleton information learning for action recognition. The proposed network is called 3 s-STNet. This network fine-tunes and fuses the spatial–temporal features learned by the three streams of ST-GCN, T-R2N and R-R2N. Among them, the T-R2N uses the TSRJI image as the input of Res2Net-101, which aims to use the Res2Net-101 to fully learn the spatial relationship between joints at multiple scales to complement for the shortcomings of the ST-GCN network in extracting single spatial features between joints. The R-R2N uses RGB image with appearance information as the input of Res2Net-101, which solves the problem that ST-GCN uses skeleton data as network input to cause unsatisfactory recognition of some actions involving object and scene

information. Compared with other action recognition methods, the 3 s-STNet proposed achieves the state-of-the-art results at present. In future research, we will further optimize the architecture of the model to improve the accuracy of action recognition and reduce the computation complexity.

Acknowledgements This work is supported partially by the National Natural Science Foundation of China under the Grant 62277009, the project of Jilin Provincial Science and Technology Department under the Grant 20180201003GX, the project of Jilin province development and reform commission under the Grant 2022C047-5. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

Author contributions This study was completed by the co-authors. SL conceived the research. The major experiments and analyses were undertaken by MF and SP. YZ and HY were responsible for data processing and drawing figures. C-CH edited and reviewed the paper. All authors have read and approved the final manuscript.

Funding This work was supported by the project of Changchun Municipal Science and Technology Bureau under the Grant 21ZY31.

Declarations

Conflict of interest Authors declare no conflicts of interest.

References

1. Yan S, Xiong Y, Lin D (2018) April. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
2. Peng W, Hong X, Chen H, Zhao G (2020) Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, No. 03, pp. 2669–2676.
3. Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N (2019) View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Trans Pattern Anal Mach Intell 41(8):1963–1978
4. Li Y, Ji B, Shi X, Zhang J, Kang B, Wang L (2020) Tea: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 909–918.
5. Sudhakaran S, Escalera S, Lanz O (2020) Gate-shift networks for video action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1102–1111.
6. Abdelbaky A,, Aly S (2020) Human action recognition using short-time motion energy template images and PCANet features. Neural Comput Appl, 1–14.
7. Li Y, Xia R, Liu X, Huang Q (2019) Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition. In: 2019 IEEE international conference on multimedia and Expo (ICME) (pp. 1066–1071). IEEE, New York.
8. Caetano C, Sena J, Brémond F, Dos Santos JA, Schwartz WR (2019) Skelemotion: a new representation of skeleton joint sequences based on motion information for 3d action recognition. In: 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1–8). IEEE, New York.
9. Caetano C, Brémond F, Schwartz WR (2019) Skeleton image representation for 3d action recognition based on tree structure and reference joints. In: 2019 32nd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI) (pp. 16–23). IEEE, New York.
10. Fang L, Wu G, Kang W et al (2019) Feature covariance matrix-based dynamic hand gesture recognition[J]. Neural Comput Appl 31(12):8533–8546
11. Zheng W, Li L, Zhang Z, Huang Y, Wang L (2019) Relational network for skeleton-based action recognition. In: 2019 IEEE International conference on multimedia and Expo (ICME) (pp. 826–831). IEEE, New York
12. Li S, Li W, Cook C, Zhu C, Gao Y (2018) Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5457–5466).
13. Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P (2019) Res2net: A new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662.
14. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725–1732).
15. Gan C, Wang N, Yang Y et al (2015) Devnet: A deep event network for multimedia event detection and evidence recounting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2568–2577.
16. Lin J, Gan C, Han S (2019) TSM: temporal shift module for efficient video understanding. In: Proceedings of the 17th IEEE International Conference on Computer Vision, Seoul, Oct 7–Nov 2, 2019. Piscataway: IEEE, 2019: 7083–7093.
17. Ji S, Xu W, Yang M, Yu K (2012) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
18. Lu X, Yao H, Zhao S, Sun X, Zhang S (2019) Action recognition with multi-scale trajectory-pooled 3D convolutional descriptors. Multimedia Tools Appl 78(1):507–523
19. Feichtenhofer C (2020) X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition, pp 203–213.
20. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199.
21. Feichtenhofer C, Fan H, Malik J et al (2019) Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6202–6211.
22. Chéron G, Laptev I, Schmid C (2015) P-cnn: Pose-based cnn features for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 3218–3226.
23. Wang H, Wang L (2017) Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 499–508).
24. Liu J, Shahroudy A, Xu D, Kot AC, Wang G (2017) Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. IEEE Trans Pattern Anal Mach Intell 40(12):3007–3021
25. Li C, Xie C, Zhang B, Han J, Zhen X, Chen J (2021) Memory attention networks for skeleton-based action recognition. IEEE Trans Neural Netw Learn Syst 33(9):4800–4814. https://doi.org/10.1109/TNNLS.2021.3061115.

26. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1227–1236).
27. Shi L, Zhang Y, Cheng J, Lu H (2019) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12026–12035).
28. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
29. Yang Z, Li Y, Yang J, Luo J (2018) Action recognition with spatio–temporal visual attention on skeleton image sequences. IEEE Trans Circuits Syst Video Technol 29(8):2405–2415
30. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3288–3297).
31. Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1010–1019).
32. Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC (2019) Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE Trans Pattern Anal Mach Intell 42(10):2684–2701
33. Xie S, Girshick R, Dollár P et al (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1492–1500.
34. Das S, Dai R, Yang D, Bremond F (2021) VPN++: Rethinking Video-Pose embeddings for understanding Activities of Daily Living. arXiv preprint arXiv:2105.08141.
35. Baradel F, Wolf C, Mille J (2017) Human action recognition: Pose-based attention draws focus to hands. In: Proceedings of the IEEE International conference on computer vision workshops (pp. 604–613).
36. Baradel F, Wolf C, Mille J (2018) Human activity recognition with pose-driven attention to rgb. In: BMVC 2018–29th British Machine Vision Conference (pp. 1–14).
37. Liu M, Yuan J (2018) Recognizing human actions as the evolution of pose estimation maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1159–1168).
38. Shi L, Zhang Y, Cheng J, Lu H (2019) Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7912–7921).
39. Das S, Dai R, Koperski M, Minciullo L, Garattoni L Bremond F, Francesca G (2019) Toyota smarthome: Real-world activities of daily living. In: Proceedings of the IEEE/CVF international conference on computer vision (pp. 833–842).
40. Das S, Chaudhary A, Bremond F, Thonnat M (2019) Where to focus on for human action recognition? In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 71–80). IEEE, New York.
41. Liu Z, Zhang H, Chen Z, Wang Z, Ouyang W (2020) Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 143–152).
42. Das S, Sharma S, Dai R, Bremond F, Thonnat M (2020) Vpn: Learning video-pose embedding for activities of daily living. In: European conference on computer vision (pp. 72–90). Springer, Cham.
43. Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, Wanli Ouyang (2020) Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings ofthe IEEE/CVF conference on computer vision and pattern recognition, p. 143–152
44. Liu J, Wang G, Hu P, Duan LY, Kot AC (2017) Global context-aware attention lstm networks for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1647–1656).
45. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6299–6308).
46. Yang H, Yan D, Zhang L, Li D, Sun Y, You S, Maybank SJ (2020) Feedback graph convolutional network for skeleton-based action recognition. arXiv preprint arXiv:2003.07564.
47. Chen Z, Li S, Yang B, Li Q, Liu H (2021) Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 2, pp. 1113–1122).
48. Cheng K, Zhang Y, He X et al (2020) Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 183–192.
49. Friji Rasha, Hassen Drira, Faten Chaieb, Hamza Kchok, Sebastian Kurtek (2021) Geometric deep neural network using rigid and non-rigid transformations for human action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 12611–12620.