Spatio-temporal Attention Network for Student Action Recognition in Classroom Teaching Videos

Yongkang Huang

Beijing University of Posts and Telecommunications

Meiyu Liang (✉ meiyu1210@bupt.edu.cn)

Beijing University of Posts and Telecommunications <https://orcid.org/0000-0003-1835-1848>

Research

Keywords: spatio-temporal residual attention neural network, multi-person classroom action recognition, transformer, object detection, object tracking

Posted Date: December 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1022972/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Spatio-temporal Attention Network for Student Action Recognition in Classroom Teaching Videos

Yongkang Huang and Meiyu Liang*

*Correspondence:
meiyu1210@bupt.edu.cn
Beijing Key Laboratory of
Intelligent Telecommunication
Software and Multimedia, School
of Computer Science, Beijing
University of Posts and
Telecommunications, Beijing,
China
Full list of author information is
available at the end of the article

Abstract

Inspired by the wide application of transformer in computer vision and its excellent ability in temporal feature learning. This paper proposes a novel and efficient spatio-temporal residual attention network for student action recognition in classroom teaching video. It first fuses 2D spatial convolution and 1D temporal convolution to study spatio-temporal feature, then combines the powerful Reformer to better study the deeper spatio-temporal characteristics with visual significance of student classroom action. Based on the spatio-temporal residual attention network, a single person action recognition model in classroom teaching video is proposed. Considering that there are often multiple students in the classroom video scene, on the basis of single person action recognition, combined with object detection and tracking technology, the association of temporal and spatial characteristics of the same student targets is established, so as to realize the multi-student action recognition in classroom video scene. The experimental results on classroom teaching video dataset and public video dataset show that the proposed model achieves higher action recognition performance than the existing excellent models and methods.

Keywords: spatio-temporal residual attention neural network; multi-person classroom action recognition; transformer; object detection; object tracking

1 Introduction

With the continuous development of artificial intelligence, intelligence has gradually entered all aspects of work and life. Concepts such as smart cities, smart offices, and smart medical care continue to emerge and develop rapidly. Smart education has gradually moved from theory to campus, which has become one of the inevitable development trends of the road of intelligent education. In traditional education, teachers need to teach dozens of students or more. The traditional teaching experience: teaching in accordance with students' aptitude and teaching in fun cannot be achieved under such a large base. Teachers do not have enough energy to give consideration to all students at the same time. Teachers can only get feedback about their teaching methods from a small number of students from occasional observations: students' expressions and action states and other listening states. Moreover, students' acceptance of what teachers teach is gradually decreasing because of the teacher's teaching is often from easy to difficult. Students may not be able to keep up with the teacher's teaching progress at a certain moment. Teachers need to observe the students' listening status frequently in order to effectively adjust their teaching progress and methods which help to achieve better teaching results.

Students' listening status includes expressions and actions. Positive expressions and actions represent that students are actively participating in classroom teaching, they can keep up with the teacher's teaching progress and think about the teaching content, but the main listening state still needs to be carried out by analyzing the students' action. Generally, students are confused or even bored with the teaching content if a student shows negative actions such as sleeping, turning heads, etc. Moreover, student's classroom action will change over time, and it is necessary to observe the changes in the student's listening action in real time to get good feedback on teaching effect in time. Corresponding analysis of students' classroom actions in the recorded teaching classroom videos after class can also help adjust teaching methods to achieve better teaching results. How to identify and analyze students' classroom actions more efficiently has become a research hotspot in intelligent education.

However, classroom teaching videos often have problems such as numerous student objects and serious occlusion. This brings great research challenges to student action recognition in classroom scenes. Therefore, it is necessary to study more robust multi-person action recognition models to automatically recognize the classroom action of all students. The current multi-person action recognition algorithms mainly include two categories, one is based on object detection and image classification, and the other is based on joint recognition and joint movement analysis. Although the former can often achieve real-time effects, it does not consider the temporal features of actions, but only classifies the object states; although the latter can learn the temporal features of the joints' position changes in the process of human movement, the recognition process of joints has a large amount of calculation, and its accuracy cannot be guaranteed. In order to solve the above problems, this paper adopted the object detection model and object tracking algorithm, which are the fastest and highly accurate at present, and combined with 3D behavior recognition algorithm to solve the problem of students' classroom action recognition in classroom scenes. Experimental results show that the proposed multi-person classroom action recognition model can achieve better performance in student classroom action recognition. In order to fully verify the proposed classroom action recognition algorithm, this paper constructed a student classroom action video dataset for model training and testing. The main contributions and innovations of this paper are as follows:

1) This paper proposes a novel and efficient spatio-temporal residual attention network for students action recognition in classroom teaching video. It first fuses 2D spatial convolution and 1D temporal convolution to learn spatio-temporal feature, then combines the powerful Reformer to better learn the spatio-temporal characteristics of student classroom action.

2) Based on the spatio-temporal residual attention network, a multi-student action recognition algorithm in classroom teaching video is proposed. It combines object detection and tracking technology to establish the correlation of the same student's spatio-temporal characteristics, and to realize the multi-student action recognition in classroom.

3) A dataset about student classroom action named Students is constructed. Experiments results shows that the proposed model achieves state-of-the-art accuracy on Students(90.87%) and HMDB51(83.46%).

2 Related Work

Currently, the multi-person action recognition algorithm is mainly divided into 2D gesture recognition and joint recognition with LSTM. There is no better solution at present. Both 2D gesture recognition and joint with LSTM methods rely on joint recognition, but it is difficult to accurately identify students' joints in complex classroom scenes. At present, the research on single-person action recognition methods has achieved better results, and Transformer has strong ability in timing information learning. How to integrate Transformer into action recognition algorithm and apply it to multi-person action recognition is also a very important research direction. So this paper uses object detection, object tracking and single person action recognition algorithm to realize student classroom action recognition in teaching classroom videos.

2.1 Object detection and object tracking

Student object detection and tracking is the first step to realize student classroom action recognition. In recent years, the object detection technology based on deep learning has achieved better performance in the field of computer vision. Existing object detection algorithms based on deep learning are divided into single-stage algorithm and two-stage algorithm [1]. The single-stage object detection algorithm is an end-to-end method, which is faster, but the accuracy is slightly lower; the two-stage object algorithm needs to first calculate the object candidate area, and then calculate object prediction on the candidate area, with higher accuracy but slower speed. The single-stage object detection algorithm is a hot research topic [2]. There are some most famous single-stage object detection algorithms. YOLO algorithm proposed by Redmon et al.[3], the latest version is YOLOv5 [1, 4], which has improved speed and accuracy. Liu et al. proposed a single-shot multi-object frame detector (SSD) algorithm [5], which introduced multi-resolution detection technology and greatly improved the accuracy of detecting single-stage detectors. The scalable and efficient object detector (EfficientDet) algorithm proposed by the Google Brain team [6], which combined the efficient network (EfficientNet) and the bidirectional feature pyramid network (BiFPN) [7], has achieved excellent object detection speed and accuracy.

After object detection is completed, it is necessary to continuously track the listening status of each student through object tracking. In the classroom scenario, if you need to analyze the classroom action of a certain student, real-time object tracking algorithms can just help. There are some object tracking algorithms that have both real-time and accuracy. Alex et al. combined the Kalman filter algorithm and the Hungarian algorithm to propose a simple online real-time tracking (SORT) algorithm [8]; Nicolaid et al. combined the SORT algorithm and the deep appearance model to propose a deep simple online real-time tracking (DeepSORT) algorithm [9, 10]; Wang et al. embedded the appearance model into the object detection model, so that only one model is needed for object tracking, and proposes the JDE algorithm [11]; Zhang et al. proposed a network FairMOT [12] that combines object detection task and Re-ID task to achieve object tracking.

2.2 Action recognition

Action recognition is currently a research hotspot in artificial intelligence, which includes three basic steps: normalization of images, extraction of action feature and action classification. The main difficulty in action recognition is that it deals with one more time dimension than image classification. And it works on videos, which has the problems of multiple scales, multiple objects, and cameras moves in an open environment. Traditional action recognition algorithms focused on capturing human contours or body information for action recognition in the early days.

With the successful application of deep learning methods in multi-object action recognition, some video action recognition methods based on deep learning have been proposed one after another. Du et al. used a three-dimensional convolutional network to extract the spatio-temporal characteristics of actions, and proposed a 3D convolutional network (C3D) for action recognition [13], which can recognize actions more accurately. Tran et al. [14] combined three-dimensional convolution and dual-stream method [15] to identify actions, and proposes an inflated convolutional network (I3D) algorithm for action recognition. Tran et al. converted three-dimensional convolution into two-dimensional convolution and one-dimensional time series convolution to extract the spatio-temporal characteristics of actions, and proposed a residual 2+1D network (R(2+1)D) [16], which can be more efficient in action recognition. Kalfaoglu et al. replaced the global average pooling layer in R(2+1)D with the BERT model [17], which can better learn the salient features of actions, and proposed the R(2+1)D-BERT model[18]. These methods learn the spatio-temporal characteristics of actions and recognize them well, but they cannot be directly applied to multi-person action recognition.

And many of the multi-person action recognition algorithms are based on the real-time multi-person 2D pose estimation (OpenPose) algorithm [19, 20]. OpenPose can detect the position of the human joints to estimate the person's posture. Song et al. [21] used OpenPose to extract the position of the human joints, and then used the SORT algorithm to track the object, calculate the offset position of the joints, and get the human action category. Chen et al. [22] used DeepSORT to track the object, and then learn the position information of the joints through the deep neural network to obtain the action type. The first two methods do not obtain the time feature of the action. Although the latter method can learn the time sequence characteristics of the joint position changes, it cannot fully express the time characteristics of the action.

2.3 Transformer

Transformer is a network widely used in the field of natural language processing (NLP), which mainly composed of self-attention mechanism. The BERT proposed by Google has caused a sensation in the academic and industrial circles in the field of NLP, and has raised the research enthusiasm of transformer to a new height. However, the computational complexity of transformer is very high, so Kitaev et al. [23] proposed an efficient transformer—Reformer. Reformer combined with local sensitive hashing and reversible layer that enables transformer to be trained on a single GPU and to have good performance. More and more transformer-based models have been proposed to improve various visual tasks due to its excellent performance[24].

Girdhar et al. [25] used a variant of the transformer architecture to aggregate contextual cues related to a specific person in the video. They prove the usefulness of this contextual information for action classification and location. Neimark et al. [26] proposed the Video Transformer Network (VTN). The network first uses 2D CNN to obtain frame-level features, and applies a transformer encoder (Longformer [27]) to learn the temporal relationship. Experiments conducted on the dataset Kinetics-400 [28] show the competitive performance of VTN. Human action recognition based on skeleton requires the model to understand the relationship between different joints of the human in a given frame, and the relationship between different frames in the video. Plizzari et al. [29] proposed a dual-stream transformer network to simulate this relationship. The dual-stream spatio-temporal transformer network implements the most advanced result on the NTU-RGB+D 60[30] and NTU-RGB+D 120[31] datasets.

At present, though there is no good solution for online multi-person action recognition, the research results of single-person action recognition are already excellent. In view of the excellent effect of single-person action recognition algorithm R(2+1)D, Reformer has excellent performance in the study of space-time characteristics, and the experimental results show that the combination of R(2+1)D and Reformer can achieve better action recognition effect. So the main method adopted in this paper is to use R(2+1)D-Reformer to realize single-person action recognition, and applies it to multi-person classroom teaching scenarios. In conclusion, this paper combines object detection, object tracking and single-person action recognition algorithms to solve student action recognition task in classroom teaching scenarios.

3 Method

This paper proposes a spatio-temporal residual attention network for students action recognition in classroom teaching video. It first fuses 2D spatial convolution and 1D temporal convolution to learn spatio-temporal feature, then combines the powerful Reformer to better learn the spatio-temporal characteristics of student classroom action. Based on the spatio-temporal residual attention network, a single person action recognition model and a multi-student action recognition algorithm in classroom teaching video is proposed. Firstly, it uses a object detection algorithm to learn the features of the classroom scene, and calculates the student object image. Then it uses a object tracking algorithm to calculate the student appearance features and match the previous features to obtain the student's continuous action image. Then the proposed spatio-temporal residual attention network is used to extract spatio-temporal characteristics of students actiona. Through the classification of students' actions characteristics, the recognition of students' classroom action is realized. The flow chart of the multi-person classroom action recognition algorithm based on proposed spatio-temporal residual attention network is shown in Figure 1.

3.1 Real-time student object detection and tracking algorithm

In order to identify the action of each student in the classroom teaching video, it is necessary to detect and track the student target in the video first, and extract the action spatio-temporal characteristics of each student. Detecting and locating students in classroom teaching videos is the key to realize multi-student action

recognition. This paper uses YOLOV5 object detection algorithm to achieve real-time detection of student in classroom teaching videos. The network structure of YOLOV5 includes a deep convolutional neural network(Backbone) that extracts object features, and a feature pyramid network(FPN) for feature fusion and a convolutional network for student detection.

After all student images are extracted by target detection algorithm, all student images are input into a simple appearance embedding model to obtain the appearance characteristics of each student. Next, Kalman filter algorithm is used to predict the appearance of each student in the next frame. After detecting the size and position of the student object in the next frame of image, the Hungarian algorithm is used to match and associate each student object to achieve classroom student object tracking. The classroom student object detection and tracking algorithm is shown in Figure 2.

3.2 Spatio-temporal residual attention network for students action characteristics learning

This paper proposes a spatio-temporal residual attention network to learn student classroom action feature. Inspired by R(2+1)D network, it extracts spatial features

of students classroom action by first using two-dimensional convolution and then uses one-dimensional convolution to extract temporal features. Then, the Reformer based on attention mechanism is used to learn the significance spatio-temporal characteristics. Compared with the traditional transformer, it can achieve better recognition accuracy with less memory consume. The network architecture of the student classroom action recognition model based on spatio-temporal residual attention network is shown in Figure 3.

R(2+1)D convolution is based on the premise that 2D convolution of a single frame video can still obtain the state-of-art results close to 3D convolution. It is a mixed convolution with residual structure. R(2+1)D network can learn the spatio-temporal characteristics of students' classroom action. It adopts the ResNet3D50 network structure, but the convolution module in each module is replaced by a 3-dimensional convolution with multiple 2-dimensional spatial convolution and a 1-dimensional time series convolution. In order to be consistent with the hyperparameters of 3D convolution, the specific calculation formula of 2D spatial convolution is shown in formula (1):

$$M_i = \lfloor \frac{td^2 N_{i-1} N_i}{d^2 N_{i-1} + t N_i} \rfloor \quad (1)$$

Where t represents the length of time, d represents the dimension of the two-dimensional spatial convolution, N_i represents the parameter of the i_{th} two-dimensional convolution, and M represents the output feature. Decomposing the 3D convolution into multiple 2D convolutions and a 1D convolution can adding a linear operation between the spatial feature and the time feature, which is easier to optimize, and smaller the training error.

In order to better learn the high level characteristics of students' classroom actions, this paper adopts a combination of R(2+1)D and Reformer to further study the high level characteristics of students' classroom actions. Reformer uses Local Sensitive Hash (LSH) instead of the general attention mechanism, which can achieve better results. The standard attention is the dot product of scaling, as shown in formula (2):

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

Where Q and K represent the query and key value of dimension d_k and V represents the value matrix. Locally sensitive hashing is to divide the hash code into buckets, and only needs to calculate the marked attention value in the same bucket, which has better sensitivity. At the same time, in order to reduce memory usage, Reformer adopts the idea of reversible layer (RevNet). The conventional residual network has a single input and single output: , while in RevNet there are pairs of input/output: , and follow the formula (3) (4):

$$y_1 = x_1 + F(x_2) \quad (3)$$

$$y_2 = x_2 + G(y_1) \quad (4)$$

Where F represents the attention layer (LSH) and G represents the feedforward layer (FFN), x_1, x_2 represent input features, y_1, y_2 represent output results for above layers. In this way, the parameters of each layer can be derived from the parameters of the current layer, so the parameters only need to be stored once instead of multiple times, which can greatly reduce the memory occupied. In addition, the latitude of the hidden layer in the middle of FFN is usually large. The general practice is to directly calculate all the results, but the input of FFN is independent. In order to reduce memory use, split calculation can be carried out to exchange time consumption for space, like formula (5), (6) and (7):

$$y_2 = x_2 + FFN(y_1) \quad (5)$$

$$y_2 = [y_2^{(1)}; y_2^{(2)}; \dots; y_2^{(c)}] \quad (6)$$

$$y_2 = [x_2^{(1)} + FFN(y_2^{(1)}); x_2^{(2)} + FFN(y_2^{(2)}); \dots; x_2^{(c)} + FFN(y_2^{(c)})] \quad (7)$$

Where c represents the total splitted blocks in FFN, $x_2^{(i)}, y_2^{(i)} (i = 0, 1 \dots c)$ represent the calculation in each block.

3.3 Students action recognition based on spatio-temporal residual attention network

Based on the characteristics of students' action obtained from the proposed spatio-temporal residual attention network, a multi-students action recognition algorithm based on the constructed spatio-temporal residual attention network was proposed. The specific algorithm is shown in Algorithm 1.

Algorithm 1 Student classroom action recognition algorithm based on spatio-temporal residual attention network

Input: classroom teaching video
Output: Classroom action category of all students
Init: object detection model D , object tracking model T and action recognition model A based on spatio-temporal residual attention network
Init: student set $M(ID, S)$, each student S has a cache C storing state images and action v

```

1: for frame in video do
2:    $boxes = D(frame)$  /* Obtain all student object images through student object detection model  $D$  */
3:    $IDs = T(boxes)$  /* Update student image assignment ID through student goal tracking model  $T$  */
4:   for ID in IDs do
5:     if ID not in  $M$  then
6:        $M.add(ID, S)$  /* Add the non-existent student  $S$  to the student set  $M$  */
7:     end if
8:      $S.C.add(frame[box])$  /* Store the corresponding student object image in the corresponding student's cache  $C$  */
9:     if  $S.C.size = 16$  then
10:       $S.v = A(S.C)$  /* Identify the types of students' classroom actions through action recognition model  $A$  */
11:    end if
12:  end for
13: end for

```

4 Experimental Results and Discussions

4.1 Dataset

In order to verify the proposed student classroom action recognition model, this paper collected 90 classroom teaching videos and constructed a classroom action library. Since it is difficult to determine the start and end of the action by manually labeling the video, this paper use the object tracking algorithm to obtain the image stream of the action state change of a single student object, and then manually select some continuous frames from the single person's image stream as a single person's action samples. The total of samples were selected-listening: 360, standing: 200, headdown: 260; turnhead 220. The sample diagram of each action sample is shown in Figure 4. In addition, the public dataset HMDB51 is also used to verify the effectiveness of the model. The dataset HMDB51 contains 51 types of actions, each with at least 51 videos. The processing method of the HMDB51 dataset is the same as the classroom teaching video dataset.

4.2 Evaluation indices

The evaluation method adopted in this paper for object detection and object tracking is to count the number of people who are correctly identified by each algorithm and the corresponding calculation time or calculation speed. The evaluation index adopted for action recognition is accuracy, and the calculation formula is:

$$ACC = \frac{True}{ALL} \quad (8)$$

Where ALL represents the number of all samples, and $True$ represents the number of samples that are predicted to be correct.

Figure 4 Sample pictures of various action samples, (a) Sample pictures of listening, (b) Sample pictures of standing, (c) Sample picture of headdown, (d) Sample picture of turnhead

4.3 Experiment results and analysis

4.3.1 Student object detection experiments and analysis

In order to verify the effect of object detection in classroom teaching videos, this paper compared several current object detection algorithms with excellent real-time and accuracy through experiments, including EfficientDet [6], YoloV4 [1] and YoloV5 [1, 4]. In the experiment, multiple trained models with different input sizes are used, and the same object probability threshold is 0.3. The experimental results are shown in Table 1. The true value of the object number of students in the video scene is 42.

Table 1 Comparison table of experimental results of classroom student object detection

Model	Image resolution(px*px)	Recognized persons	Cost time(ms)
EfficientDet	512*512	16	71
	640*640	23	151
	768*768	20	234
	896*896	23	477
	1024*1024	23	795
YoloV4	416*416	14	290
	608*608	20	510
YoloV5s	608*608	22	24
YoloV5m		19	50
YoloV5l		24	79
YoloV5x		26	179

YOLOV5 model has achieved great results in terms of speed and accuracy. Through the analysis of the experimental results in Table 1, it can be found that

YOLOV5x model can identify 26 students with the highest object detection accuracy. When YOLOV4 model uses 416 * 416px video frame as input, the number of people recognized is at least 14, followed by efficientdet model which uses 512 * 512px video frame as input, only 16 people are recognized, and the number of people recognized by other models is about 20, the effect difference is not obvious. In terms of detection time, YOLOV5 only takes 24 ms to process a single video frame, which can be processed in real time, and other YOLOV5 models also take less time. The slowest is when efficientdet takes 1024 * 1024 video frames as input, the processing time of a single video frame is 795ms, which completely does not meet the real-time requirements. This should be due to the differences in the network structure of the model. Therefore, the current experiment is mainly based on YOLOV5 algorithm to realize the real-time student detection. In order to verify the effect of YOLOV5 models in different teaching scenes, the YOLOV5 models are tested on the video frames of five different teaching scenes, and the comparison results are shown in Table 2.

Table 2 Comparison table of student object detection experiment results in different classroom scenarios

Frame ID	YOLOV5s	YOLOV5m	YOLOV5l	YOLOV5x
1	10	12	11	10
2	18	16	23	23
3	18	24	19	22
4	30	22	29	31
5	18	14	18	18

Yolov5s model has achieved great results in terms of speed and accuracy. YOLOV5s model According to the experimental results shown in Table 3, YOLOV5s model has the same effect as other models on video frames 1, 4 and 5, and is close to the best object detection accuracy, but has poor detection effect on video frames 2 and 3. The YOLOV5m model achieves the best detection accuracy on video frames 1 and 3, but it is similar to other models, while its detection accuracy on other video frames is close to the worst. The YOLOV5l model achieves the best detection accuracy on video frames 2 and 5, as well as other video frames. The YOLOV5x model has the best object detection accuracy on video frames 2, 4 and 5. It also works well on other video frames. However, in terms of overall object recognition effect and recognition time, the YOLOV5m model has the worst object detection effect in classroom teaching scenarios. Although the YOLOV5l and YOLOV5x models have high detection accuracy, the detection time is too long, while YOLOV5s can achieve a balance between object detection accuracy and detection time of students. Therefore, YOLOV5s model was used as a real-time object detection model for classroom students in subsequent experiments.

4.3.2 Student object tracking experiments and analysis

In order to verify the performance of student object tracking in classroom teaching video scene, this paper compares the current object tracking algorithms with good real-time performance and accuracy, including YoloV5s [1, 4] + DeepSORT [9, 10], JDE[11] and FairMOT[12]. The experimental results of different models are shown in Table 3. By observing the experimental results in Table 3, this paper can find that YoloV5s [1, 4] + DeepSORT [9, 10] object tracking algorithm has achieved good

results in terms of speed and accuracy, while JDE[11] and FairMOT[12] model can only achieve certain results in multi pedestrian scene, but the speed is also slow. To sum up, YoloV5s [1, 4] + DeepSORT [9, 10] algorithm is used to achieve student object tracking in classroom teaching video.

Table 3 Comparison of experimental results of different object tracking algorithms in different scenarios

Scene		YoloV5s-DeepSORT	Fair-MOT	JDE
Multi-walker	speed (fps)	33	11	14
	persons	24	60	55
Teaching	speed (fps)	33	13	17
	persons	19	1	4

4.3.3 Student classroom action recognition experiments and analysis

In order to verify the effects of different action recognition algorithms in classroom scenarios, the proposed model is compared with I3D[14], R(2+1)D [16], ResNet3D101[32], ResNeXt3D101[33] and the models incorporating BERT or Reformer into these models. Experimental parameters: epochs=50, batch-size=8, clip-len=16, input-size=112*112, learning rates are 0.01, 0.001, 0.0001 and 0.00001. The experimental results of different student action recognition models are shown in Table 4, where the average accuracy rate is the highest value on the test dataset in multiple trainings.

Table 4 Comparison table of the accuracy of each action recognition algorithm on the student classroom action dataset

Model name	Extra training data	ACC(%)
I3D	ImageNet	79.33
I3D-BERT		87.02
I3D-Reformer		84.13
ResNet3D50	Kinetics400	85.58
ResNet3D50-BERT		87.50
ResNet3D50-Reformer		86.06
ResNet3D101	Kinetics400	87.02
ResNet3D101-BERT		87.50
ResNet3D101-Reformer		87.98
ResNeXt3D101	Kinetics400	87.50
ResNeXt3D101-BERT		85.58
ResNeXt3D101-Reformer		88.46
R(2+1)D	Sports-1M	88.46
R(2+1)D-BERT		87.02
R(2+1)D-Reformer(ours)		90.87

By observing the experimental results in the table 4, it can be found that except for ResNeXt3D101, the accuracy of other backbone networks has been improved to varying degrees after fusing the BERT or Reformer structure. BERT and Reformer have the greatest impact on the I3D model, increasing by 7.69% and 4.8 respectively. And the impact on ResNet3D101 is the smallest, only increased by 0.48% and 0.96%, indicating that the BERT and Reformer mechanism has a certain gain for action recognition tasks. In addition, our proposed model achieves the highest accuracy 90.87% on the Student classroom action dataset, which is 2.41% higher than R(2+1)D and 3.85% higher than R(2+1)D-BERT, indicating that Reformer integrated into the R(2+1)D can indeed improve its high-dimensional time series feature learning ability and improve the performance of the it.

Further observe the training curve of BERT and Reformer added to R(2+1)D in Figure 5, and it can be found that R(2+1)D-Reformer is not only the highest

Figure 5 Parameter change curve of model training process of R(2+1)D with different Transformer, (a) Training set loss curve, (b) training set accuracy curve, (c) validation set loss curve, (d) validation set accuracy curve

accuracy, but also the decreasing speed of training loss and validation set loss. The accuracy increase rate on training set and validation are also the fastest, while the convergence rate of R(2+1)D-BERT is the slowest. It shows that BERT and Reformer are useful for learning deep features, and Reformer has faster learning speed and better performance. Although BERT can also improve the model's ability to learn deep features, it reduces the learning speed of the model, which may be caused by the higher complexity of Bert's model than that of Reformer.

In order to further verify the effect of the structure of the reformer on the whole model, this paper modifies the depth of the reformer, train the model, and carries out experiments on the test set. The experimental results are shown in Table 5.

Table 5 Accuracy comparison table of the model on the open data set HMDB51

Reformer depth	ACC(%)	Model size(MB)	Training time(min)
1	90.87	511	40.5
2	88.46	533	94.5
3	87.50	555	120

Increasing the number of Reformer layers has no good effect on R(2+1)D-Reformer. The comparison results show that increasing the depth of the Reformer does not increase accuracy, but it will increase the size of the model and significantly increase the training time. This should be caused by the architecture of transformer network and the calculation method of attention mechanism. In summary, the proposed model performs best when the depth of Reformer is 1.

Figure 6 Action recognition effect in the classroom teaching scene recognition of the proposed model, (a)example 1, (b) example 2, (c) example 3, (d) example 4

The trained spatio-temporal residual attention network is embedded in the multi-person classroom action recognition algorithm. This paper conducted experiments on the algorithm in four scenes, and the recognition effect diagram is shown in Figure 6.

In Figure a, the standing, headdown, and listening states are correctly identified. The listening state identified in Figure b are also correct. In Figure c, the standing and headdown state identified are correct, and most of the listening states are also correct, but the head turning state in the lower right corner is not recognized. In Figure d, the standing and listening states identified are correct, but the turn-head state recognized in the lower right corner is incorrect. The student has only a small part in the image may cause the wrong recognition result. In summary, the multi-person classroom action recognition algorithm proposed in this paper based on object detection and tracking and the single-person action recognition algorithm R(2+1)D-Reformer has achieved very good results. In conclusion, the proposed model has achieved good results in all four scenarios.

4.3.4 Experiment result and analysis on HMDB51

In order to verify the generalization ability of the proposed model, this paper conducted experiments on the public dataset HMDB51. There are three pre-training data sets: ImageNet, Kinetics400 and IG65M, and some models use flow data. The accuracy of the proposed model and other advanced action recognition models on HMDB51 are shown in Table 6.

The results in table 6 shows that the accuracy on HMDB51 of the proposed R(2+1)D-Reformer is 24.06% higher than the Two-Stream, 12.46% higher than the TSN, and 2.56% higher than the I3D without using Flow, indicating that the proposed R(2+1)D-Reformer has a very strong ability to learn the spatio-temporal

Table 6 Accuracy comparison table of the model on the open data set HMDB51

Model	Use Flow?	Extra Training Data	ACC(%)
Two-Stream [34]	Y	ImageNet	59.40
TSN [35]	Y	ImageNet	71.00
TSM [36]	N	Kinetics400	73.50
I3D [14]	Y	Kinetics400	80.90
R(2+1)D(32f) [16]	N	IG65M	80.54
R(2+1)D-Reformer(32f ours)	N	IG65M	83.46

characteristics of action. And without flow data, the accuracy of the R(2+1)D-Reformer is 9.94% higher than that of TSM. The accuracy of the R(2+1)D-Reformer is 2.98% higher than that of the R(2+1)D, reaching an advanced level, indicating that Reformer's ability to learn time-series features has been successfully added to R(2+1)D, and achieved good results.

5 Conclusions

This paper proposes a spatio-temporal residual attention network for students' classroom action recognition tasks in classroom video scenes. It first fuses 2D spatial convolution and 1D temporal convolution to study spatio-temporal feature, then combines the powerful Reformer to better study the spatio-temporal characteristics of student classroom action. Based on the proposed spatio-temporal residual attention network, a single person action recognition model and a multi-student action recognition algorithm in classroom teaching video was proposed. It combines object detection and tracking technology to establish the correlation of the same student's spatio-temporal characteristics, and to realize the multi-student action recognition goal in classroom. A dataset about student classroom action named Students is constructed. Experiments results shows that the proposed model achieves state-of-the-art accuracy on Students(90.87%) and HMDB51(83.46%).

Acknowledgements

The authors would like to thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

This work is partially supported by the National Natural Science Foundation of China (61877006).

Abbreviations

LSTM: Long short-term memory, YOLO: You only look once, SSD: Single shot detector, BiFPN: Bidirectional feature pyramid network, SORT: Simple online and realtime tracking, JDE: Jointly learns the detector and embedding model, FairMOT: Fair Multi-object tracking, NLP: Natural language processing, VTN: Video transformer network, CNN: Convolution neural network, FPN: feature pyramid network

Availability of data and materials

Some data can be shared on request.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors have equivalent contributions. All authors read and approved the final manuscript.

Author details

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China.

References

- Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
- Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055 (2019)

3. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
4. Itlalytics: Yolov5 (2020). <https://github.com/ultralytics/yolov5> Accessed Accessed 20 Feb 2020
5. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37 (2016). Springer
6. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)
7. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR
8. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3464–3468 (2016). IEEE
9. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649 (2017). IEEE
10. Wojke, N., Bewley, A.: Deep cosine metric learning for person re-identification. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 748–756 (2018). IEEE
11. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, pp. 107–122 (2020). Springer
12. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 1–19 (2021)
13. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
14. Carreira, J., Zisserman, A.: 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 (2017)
15. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199 (2014)
16. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
17. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
18. Kalfaoglu, M.E., Kalkan, S., Alatan, A.A.: Late temporal modeling in 3d cnn architectures with bert for action recognition. In: European Conference on Computer Vision, pp. 731–747 (2020). Springer
19. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)
20. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1145–1153 (2017)
21. TianzhongSong: Real-Time-Action-Recognition (2017). <https://github.com/TianzhongSong/Real-Time-Action-Recognition> Accessed Accessed 20 Feb 2020
22. LZQthePlane: Online-Realtime-Action-Recognition-based-on-OpenPose (2018). <https://github.com/LZQthePlane/Online-Realtime-Action-Recognition-based-on-OpenPose> Accessed Accessed 20 Feb 2020
23. Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451 (2020)
24. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on visual transformer. arXiv preprint arXiv:2012.12556 (2020)
25. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253 (2019)
26. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. arXiv preprint arXiv:2102.00719 (2021)
27. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
28. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
29. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: International Conference on Pattern Recognition, pp. 694–701 (2021). Springer
30. Shahroudy, A., Liu, J., Ng, T.-T., Wang, G.: 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 (2016)
31. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
32. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
33. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
34. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199 (2014)
35. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2740–2755 (2018)

36. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7083–7093 (2019)