EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

A spatio-temporal attention fusion model for students behaviour recognition

Xiaoli Wang[1,*]

[1]School of Continuing Education, SanMenXia College of Social Administration, SanMenXia, 472000, China

Abstract

Student behavior analysis can reflect students' learning situation in real time, which provides an important basis for optimizing classroom teaching strategies and improving teaching methods. It is an important task for smart classroom to explore how to use big data to detect and recognize students behavior. Traditional recognition methods have some defects, such as low efficiency, edge blur, time-consuming, etc. In this paper, we propose a new students behaviour recognition method based on spatio-temporal attention fusion model. It makes full use of key spatio-temporal information of video, the problem of spatio-temporal information redundancy is solved. Firstly, the channel attention mechanism is introduced into the spatio-temporal network, and the channel information is calibrated by modeling the dependency relationship between feature channels. It can improve the expression ability of features. Secondly, a time attention model based on convolutional neural network (CNN) is proposed, which uses fewer parameters to learn the attention score of each frame, focusing on the frames with obvious behaviour amplitude. Meanwhile, a multi-spatial attention model is presented to calculate the attention score of each position in each frame from different angles, extract several saliency areas of behaviour, and fuse the spatio-temporal features to further enhance the feature representation of video. Finally, the fused features are input into the classification network, and the behaviour recognition results are obtained by combining the two output streams according to different weights. Experiment results on HMDB51, UCF101 datasets and eight typical classroom behaviors of students show that the proposed method can effectively recognize the behaviours in videos. The accuracy of HMDB51 is higher than 90%, that of UCF101 and real data are higher than 90%.

Keywords: student behavior, spatio-temporal attention, channel information, multi-spatial attention, CNN.

Received on 10 August 2021, accepted on 26 August 2021, published on 03 September 2021

doi:10.4108/eai.3-9-2021.170905

[*]Corresponding author. Email: 910675024@qq.com

1. Introduction

Artificial intelligence technology and big data technology have promoted the transformation of modern education system [1,2]. Adaptive personalized learning driven by artificial intelligence technology is the most potential application scenario in the field of education. As the main place of classroom teaching in colleges and universities, multimedia classroom has been gradually upgraded to smart classroom. Classroom is also the main battlefield of "golden course" construction. Teachers play a decisive role in the construction of "golden course". How to do fusion innovation, how to effectively improve the quality of "golden course" construction, and how to effectively analyze and evaluate classroom dynamic generative teaching data have been widely concerned by education experts and front-line teachers. At present, the research focus is on the theoretical analysis, technical application and value discussion of the dynamically generated content. There are few researches on the teaching and learning data recording, data analysis and teaching application of the dynamically generated content. However, the key points and difficulty of these researches lie in the automatic detection and recognition of students' classroom behavior.

Behaviour recognition [3] has been widely used in many fields, such as video surveillance, smart home, video retrieval, intelligent human-computer interaction, etc. Video has the characteristics of complex environment, large transformation range of visual angle and human behaviour, which makes the feature representation of video have a lot of redundant information in spatio-temporal. Therefore, it is very important for behaviour recognition to effectively utilize the information of key areas on the frames with obvious behaviour amplitude in the video.

Behaviour recognition methods in the video can be divided into traditional methods [4,5] and deep learning-based methods [6,7]. Traditional methods have made some progress in the field of behaviour recognition, but they rely heavily on artificial feature design, and the generalization ability of the algorithm is insufficient. Deep learning-based methods can automatically learn the features of videos for classification, especially, the dual-stream method [8] can effectively combine the spatio-temporal information in videos and has relatively better performance. Dai et al. [9] proposed the dual-stream model for the first time, which input single-frame image and multi-frame density optical flow field image into spatial flow and temporal flow respectively. Then it fused and classified the features of the two streams. Wang et al. [10] proposed temporal piecewise network, using sparse sampling and video supervision strategies to further improve the recognition accuracy. However, the dual-stream method can not effectively utilize the key spatio-temporal information of video, and it ignores the information difference of different channels when extracting video features. In order to obtain the information of saliency regions in the video, references [11,12] used object detection or posture estimation to extract multiple key regions or body parts in the video, and then input them into the network for behaviour recognition. However, object detection or posture estimation in advance will increase the overall calculation cost. Moreover, the results of detection and estimation can affect the performance of recognition.

The behaviour recognition method based on attention mechanism [13] can automatically learn the key information in the video. Hu et al. [14] designed a channel attention network to model features from channels to highlight key channel information. Sharma et al. [15] proposed the spatial attention model to highlight the saliency areas in each frame. Du et al. [16] used the temporal attention model designed by recurrent neural network (RNN) to assign corresponding weights to different frames, which could effectively utilize the key frames of the video. Yang et al. [17] used bidirectional LSTM to design a spatio-temporal attention model. The above methods have the following deficiencies:

a) The time attention model designed by RNN or LSTM has many parameters. RNN has a fixed serial structure, so video frames must be processed in accordance with the sequence of time, and the recognition efficiency is low.

b) When extracting spatial saliency information, it will lead to the problem of inaccurate information of the extracted regions using only one spatial attention model to extract multiple behaviour regions of a frame.

To solve the above problems, this paper proposes a new students behaviour recognition method based on spatio-temporal attention fusion model. The main contributions of this paper are as follows.

1) The channel attention is integrated into the spatio-temporal network, and the channel information of the features is recalibrated while considering the spatio-temporal features, which enhances the expression ability of the features.

2) Attention model based on CNN is proposed to focus on the frame with a strong understanding on the temporal domain. Compared with the temporal attention of RNN model, this model calculates the attention score of each frame in the temporal dimension of the video. The model has fewer parameters and the calculation cost is small. It can realize the parallel operation of multiple frames and improve the overall operation efficiency.

3) A multi-spatial attention model is proposed to learn the weight of each frame from different angles by using multiple models to obtain multiple discriminant behaviour regions, which reduces the interference of background information.

4) The temporal and spatial features are fused to further enhance the feature representation of the video. Experiment results on UCF101, HMDB51 datasets and eight typical classroom behaviors of students show that the proposed model is an end-to-end and efficient behaviour recognition model.

2. Spatio-temporal attention mechanism for behaviour recognition

The video can be regarded as a combination of spatial and temporal. In spatial, RGB images contain the appearance information about the scenes and objects. In temporal, the optical flow image includes the behaviour information of the object. In this paper, the appearance flow with RGB image and the behaviour flow with optical flow image are used as the design basis. A new behaviour recognition model is proposed to enhance the feature representation, distinguish the features of different channels, and focus on the multiple saliency areas of behaviour in the frames with strong discriminant power, so as to realize the behaviour recognition. The overall structure of the proposed recognition model is shown in figure 1. In order to obtain appropriate input fragments, the new model performs sparse sampling on the video. The implementation method is as follows: dividing the video into N segments at equal intervals, sampling one frame randomly for each segment, and inputting the RGB image and optical flow image into the spatio-temporal network.

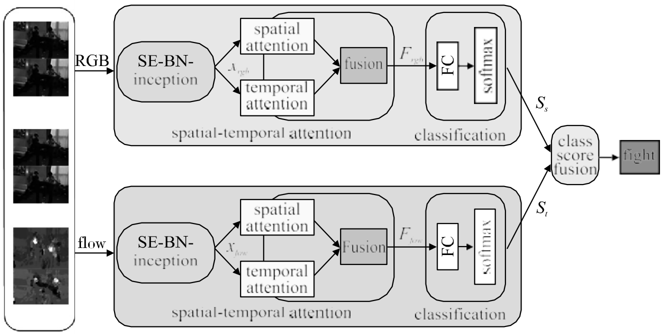


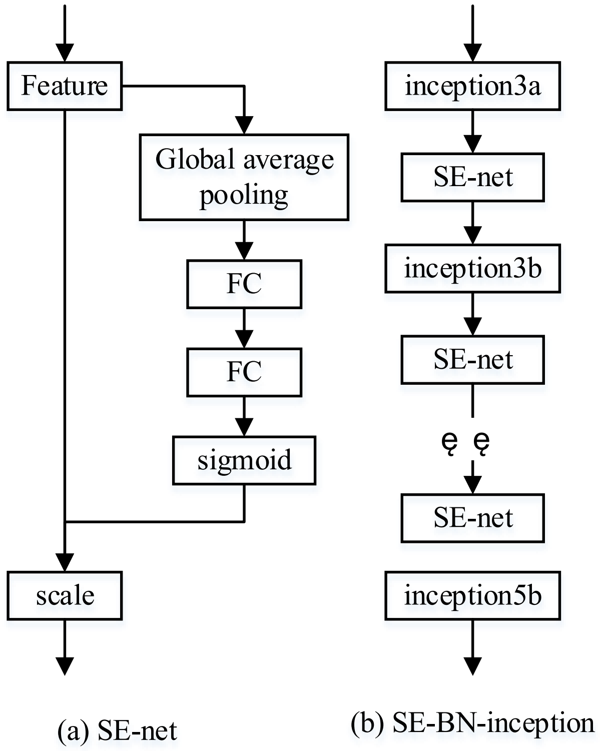
Figure 1. Structure of the proposed spatio-temporal network

The proposed method includes appearance flow and behaviour flow, and each flow network contains three modules: 1) the SE-BN-Inception module can distinguish the difference of different channel presentation features, and extract video features with strong expression ability. The appearance flow output is x_{rgb}, and the temporal flow output is x_{flow}. 2) The spatio-temporal attention module can further enhance the feature representation of the video, and highlight the frames with strong recognition and several saliency behaviour regions of the frames in spatial and temporal respectively through the temporal attention model and the multi-spatial attention model. 3) The classification module consists of a FC layer and a Sofimax function. The spatio-temporal feature F_{rgb} and F_{flow} of the two streams are input into the classification module respectively to obtain the output S_s of the appearance stream and the output S_t of the behaviour stream. The final behaviour recognition result is obtained by fusing the outputs of the two streams according to different weights.

3. SE-BN-Inception module

Multi-channel feature vectors are generated when features of video frames are extracted using convolutional networks. Each channel of the vector describes the current frame in a specific way, and different channels represent information of varying importance. However, the previous deep learning-based feature extraction methods ignored the differences of different channels, resulting in poor feature representation capability. The channel attention mechanism can learn the importance of each feature channel, increase the channel features that are useful for current recognition according to the importance, and suppress the channel features with weak recognition power. This paper introduces the channel attention implementation network SE-net (Squeeze-and-excitation network) to the BN-inception [18]. The SE-BN-Inception module is obtained to calibrate the information of different channels and enhance the expression ability of video features.

SE-net is shown in figure 2 (a). Firstly, the input features are pooled globally along the channel dimension. The dependencies between the channels are then modeled through the two fully connection layers. The first fully connection layer reduces the input channel dimension by 1/16 to reduce computation. And then it increases the nonlinearity by activating the ReLU function. The second fully connection layer returns the channel to its original dimension. The normalized weights are obtained by a sigmoid function. Finally, the weight is weighted to the features of each channel through feature redirection operation. As shown in figure 2(b), SE-BN-Inception consists of nine Inception operations. The SE-net is added after each inception. Because the output of the fully connection layer is not sensitive enough to space and position, the output of the convolution layer preserves the spatial structure of the image to a certain extent, so BN-Inception is retained to the last convolution layer.



(a) SE-net (b) SE-BN-inception

Figure 2. Structure of SE-net and SE-BN-inception

4. Spatio-temporal attention module

The spatio-temporal attention module is composed of CNN-based temporal attention model [19], multi-spatial attention model and the fusion of spatio-temporal features. The temporal attention model and the multi-spatial attention model focus on key frames and multiple saliency behaviour regions from the temporal and spatial dimensions of the video, respectively. The fusion of spatio-temporal features can effectively combine the extracted key spatio-temporal information, further enhance the feature representation of video, and improve the accuracy of behaviour recognition.

4.1. CNN-based temporal attention model

Behaviour is a process of constant change. Different frames in a video have different contributions to behaviour recognition, so the frames with rich information and obvious behaviour changes should be selected for classification. The temporal attention model can give more attention to the key frame. However, the previous temporal attention model is designed and implemented based on RNN, which has many network parameters, complex structures and it cannot be parallelized over time. In order to solve this problem, this paper proposes a temporal attention model based on CNN, which uses CNN to generate the attention score of each frame. The attention score is used to determine the importance of each frame in the video relative to the behaviour recognition. It selectively focuses on the key frames. The video feature representation is further enhanced in time dimension. The temporal attention model designed in this paper not only has fewer parameters and a simple structure, but also can calculate the attention score of all frames in parallel, it makes full use of the advantages of GPU hardware. The CNN-based temporal attention model is shown in figure 3.

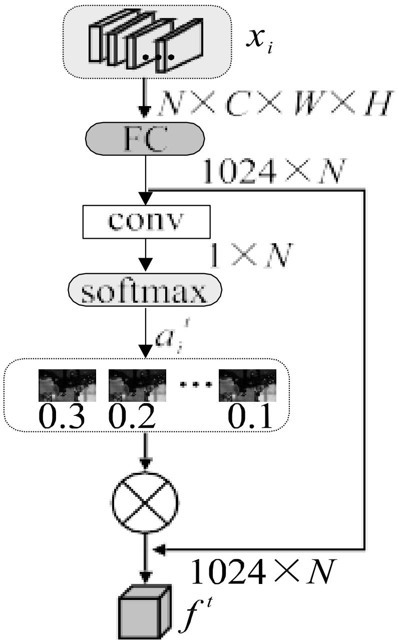


Figure 3. Temporal attention model based on CNN

After processed by SE-BE-Inception, the feature is $X=(x_1,\cdots,x_i,\cdots,x_N)$, $X\in R^{N\times C\times W\times H}$. N represents the selected frame number of the video. C represents the feature dimension degree. $W\times H$ represents the number of grid cells of the feature map. For the feature vector x_i of i-th frame of the video, it is first linearly mapped through the full connection layer, and the mapped feature is $\hat{x}_i$. The linear mapping of the same video frame uses the same parameters as shown in equation (1).

$$\hat{x}_i = w_1 x_i + b_1, i = 1,2,\cdots,N \qquad (1)$$

Where w_1 and b_1 are the learning parameters in the model. The map feature of the whole video is

$\hat{X} = (\hat{x}_1, \cdots, \hat{x}_i, \cdots, \hat{x}_N)$, $\hat{X} \in R^{N \times D}$, D=256. The video feature dimension is changed to 1×N through a convolution layer with size of 1×1. It uses the softmax function along the time dimension of the video frame to get the time attention score α_i^t of each frame in the video:

$$\alpha_i^t = \frac{\exp(conv(\hat{x}_i))}{\sum_{i=1}^{N} \exp(conv(\hat{x}_i))} \qquad (2)$$

Where *conv* represents the convolution operation. α_i^t represents the contribution of i-th frame to the recognition. After the attention score α_i^t of i-th frame is obtained, the time feature of i-th frame is obtained by multiplying it with features. The time features of all frames is summed to get the temporal feature f^t of the whole video.

$$f^t = \sum_{i=1}^{N} \alpha_i^t x_i, i = 1, 2, \cdots, N \qquad (3)$$

Where $f^t = R^{1 \times D}$, it considers the importance of each selected frame in the video.

4.2. Multi-spatial attention model

Video consists of sequential images, and each frame can be divided into regions with saliency behaviour and other regions in spatial. For behaviour recognition videos, the saliency behaviour areas are usually the moving parts of the human body and the position of the moving objects, such as the behaviour of drinking water. The behaviour can be accurately recognized by using the features of the arm, head area and the cup. Therefore, the focus should be placed on areas with significant behaviour in each frame. Generally, object detection [20], posture estimation [21] and other methods are used to extract the information of key regions for behaviour recognition, which results in large workload and complex implementation.

Spatial attention mechanism can solve the above problems. However, in references [22,23], only one spatial attention model is used to extract information of different saliency regions. Some of the extracted saliency regions are inaccurate. In order to accurately extract the spatial information of different regions of the frame that interact with the behaviour, this paper proposes a multi-spatial attention model, the specific structure is shown in figure 4.

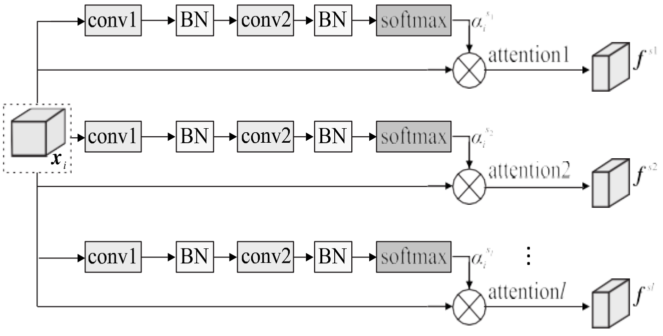


Figure 4. Multi-spatial attention model

The multi-spatial attention model does not decompose the input image in space according to the grid size of the feature map, but it extracts the spatial information of frames from multiple angles, calculates the attention score of each position in each frame, and then finds different significant areas of behaviour. This learning method can reduce the interference of irrelevant information such as background, alleviate the problems caused by human gesture changes in the video, and further enhance the feature representation of the video in space. The model number of the spatial attention represents the saliency area number of learned behaviour, and the value of the model number of spatial attention is determined through experiments. In this paper, multiple spatial attention model is used to extract saliency behaviour areas of frames. Each model is mainly composed of two convolution layers and a Softmax function. The Softmax loss is a flexible learning objective with adjustable inter-class angular margin constraint. It presents a learning task of adjustable difficulty where the difficulty gradually increases as the required margin becomes larger. For the j-th spatial attention model, X is first reduced to N×F×W×H (F=256) through a 1×1 convolution layer and tanh activation function to reduce the calculation cost. And then feature $c_j^{s_j}$ is obtained through the second convolution layer. The implementation is shown in formula (4). The BN (Batch normalize) operation is added after each convolution layer. The introduction of BN operation can solve the problem of covariance deviation and make the training more stable. The specific implementation is shown in formula (5).

$$c_j^{s_j} = BN(w_3(\tanh(BN(w_2 x_i + b_2))) + b_3) \qquad (4)$$

$$v^i = \frac{u^i - m}{\sqrt{\text{var}}} \times \alpha + \beta \qquad (5)$$

Where w_2, w_3, b_2, b_3 are the learning parameters in the network. The size of the convolution kernel of the second convolution layer is 5×5 and the convolution step is 1. $c_j^{s_j} \in R^{T \times l \times W \times H}$, l denotes the model number of the spatial attention. In formula (5), v^i and u^i are the input and output signals. α and β are trainable parameters, and m and var represent mean and variance.

The feature $c_j^{s_j}$ after two convolution layers is input into the Softmax function to calculate the probability score $\alpha_j^{s_j,k}$ of each spatial region in i-th frame.

$$\alpha_j^{s_j,k} = \frac{\exp c_i^{s_j,k}}{\sum_{k=1}^{W \times H} \exp c_i^{s_j,k}} \tag{6}$$

The weighted spatial feature is obtained by multiplying elements of $\alpha_j^{s_j,k}$ with each mapping feature. Since l spatial attention is used, l spatial features can be extracted per frame. The j-th spatial feature in selected frame of each video is summed to obtain the j-th spatial feature f^{s_j} of the whole video.

$$f^{s_j} = \sum_{i=1}^{N} \sum_{k=1}^{W \times H} \alpha_i^{s_j,k} x_i^k \tag{7}$$

4.3. Spatio-temporal feature fusion

Spatio-temporal feature fusion is used to judge the categories of human behaviours by combining the temporal and spatial features extracted from video. The fusion of spatio-temporal features can represent the change information of key frame's saliency area of behaviour, which further enhances the expression ability of features and carries out more accurate recognition of behaviour. For example, when playing golf, frames with obvious swing behaviour will get more attention through the temporal attention model. Through spatial attention model, the arm, golf club, ball and other key areas are extracted. The spatial and temporal features can be fused to focus on several saliency motion areas on the frame with obvious swing action, so as to better recognize the behaviour. The fusion of features is shown in figure 5.

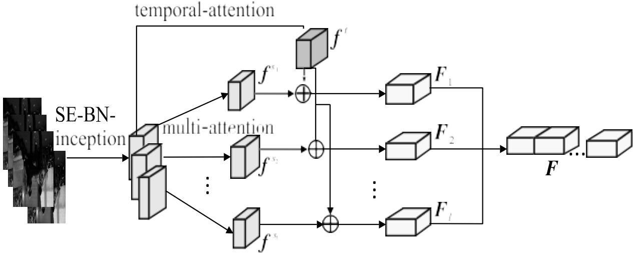


Figure 5. Fusion of spatio-temporal feature.

l spatial features f^{s_j} and one temporal feature are obtained for each video. First, it maps each spatial feature to a temporal feature. That is, l features F_l are obtained by adding the spatial feature f^{s_j} of the video and the temporal feature f^t of the video respectively. Then it connects these l features to get the spatio-temporal feature F of the video:

$$F_l = f^{s_j} + f^t \tag{8}$$

$$F = concate(F_1, F_2, \cdots, F_l) \tag{9}$$

Where *concate* denotes the connect operation.

5. Experimental results and analysis

5.1. Experimental data sets and evaluation criteria

The data sets used in this paper are two publicly available video data sets UCF101 and HMDB51 [24]. Then we also select real classroom behaviours. The UCF101 data set contains 101 behaviours and 13320 videos. The data set has a strong diversity in behaviour acquisition, including camera motion, object appearance motion, attitude change and background change. The movement category is divided into five groups: human-object interaction, body movement, person-to-person interaction, playing musical instruments and sports. The data set has problems such as large intra-class differences and small inter-class differences. HMDB51 data set contains 6676 videos and 51 types of actions. The video samples are mainly from public data such as movies, Youlube and Google video, but many videos are with poor quality. Therefore, it is challenging to perform behaviour recognition on the two data sets. For the two data sets, this paper adopts the official division method, that is, each data set is divided into three splits, 70% of the videos are training sets and 30% are testing sets.

In this paper, 60 students majoring in software engineering in 2020 from one university are selected as the research objects. The involved two courses are "Fundamentals of Programming" and "Data Structure". Two complete lectures are recorded for each course. The analysis algorithm in this paper is based on the video data as the data input object, and the camera adopts the television broadcast system (PAL), which is 25f/s (frames per second). There are four classroom teaching videos, each of which lasts 50 minutes. One classroom teaching video of each course is divided into two training sets, and the rest are used as testing sets. In order to facilitate the comparison between different taught courses in the same class, we separate the annotated data of different courses in the testing set independently and form two independent test sets, named test set A and test set B respectively.

According to the various manifestations of students classroom behavior, we focus on the basic behavior categories that can reflect students' basic states and constitute complex learning activities. In this study, eight classroom behaviors are recognized and analyzed including concentration, interaction, bowing their heads, playing with mobile phones, sleeping, reading, writing and mind wandering. The performance of the proposed algorithm is evaluated. Therefore, it is necessary to annotate the training sets and testing sets, and manually complete the coding of four videos.

In this paper, top-1 recognition accuracy is adopted as the evaluation standard. The recognition accuracy of each data set is obtained by weighted average of the action recognition accuracy of its three splits.

5.2. Experiments set

This experiment is performed on the GPU with PyTorch. the used backbone in this article is BN-Inception. BN-inception model is an upgraded version of GoogleNet model, which has a good balance between accuracy and efficiency. The network is initialized using model parameters pre-trained on the ImageNet dataset. In order to keep the optical flow data consistent with RGB data, this paper first adopts TV-L1 algorithm to obtain optical flow data, and then quantifies optical flow data to [0,255] through linear transformation.

a) Training stage. Firstly, the size of the input frame is adjusted to 240×320, and then the size of the clipping area is adjusted to 224×224 by using fixed corner clipping and horizontal flip. It adds the Dropout layer before the full connection layer of the classification network. The dropout values are set to 0.8 and 0.7 for appearance and behaviour flow, respectively. The parameters are optimized by small batch random gradient descent algorithm, and the batch size is 32. The weight attenuation coefficient is set to 0.0005. The momentum is set to 0.9. The appearance flow starts with a learning rate of 0.001. After 30 epochs and 60 epochs, it is reduced to 1/10 of the original epoch, and a total of 80 epochs are trained. The initial learning rate of the behaviour flow is 0.001, which is reduced to 1/10 after 190 epochs and 300 epochs, respectively. 340 epochs are trained.

b) Test stage. 25 frames are selected from each sample using mean sampling. For each frame image, data is enhanced by cropping and flipping, and 10 test samples are obtained. Classification results are obtained by averaging the output category probability of 10 samples.

5.3. Experimental Analysis

In this paper, the performance of behaviour recognition under different segments of video, different spatial attention models and different fusion weights are compared. Then the performance of behaviour recognition with channel attention network is analyzed experimentally. Finally, the effectiveness of the proposed method is analyzed by comparing the proposed method with the state-of-the-art methods.

Performance analysis of behaviour recognition in different video segments

In this paper, the sparse sampling method is used to sample the frames in the video and take them as the input data of the network. To analyze the influence of different video segment number on behaviour recognition performance, this paper carries out a comparative experiment on the first split of HMDB51 data set. 3, 4, 5 and 6 segments are sparsely sampled from the video for behaviour recognition, and the experimental results obtained on the appearance flow are shown in figure 6. The experimental results show that the recognition accuracy increases with the increase of the number of video segments. When the number of video segments is 6, the network has the highest recognition accuracy, because the network can learn more information from an increasing number of samples. As can be seen from figure 6, when the number of video segments is greater than 5, the rising trend of recognition accuracy gradually slows down with the increase of the number of segments. Moreover, due to the limited computer video memory, more segments cannot be tested. In this paper, each video is divided into six segments for subsequent experiments.

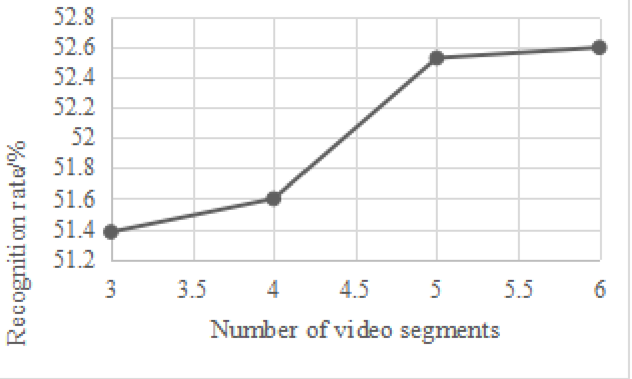


Figure 6. Comparison of recognition accuracy with different video segment number

Performance analysis of behaviour recognition under different spatial attention models

The multi-spatial attention model proposed in this paper can extract multiple saliency behaviour regions for behaviour recognition. With the increase of the spatial attention model number, the extracted saliency areas of behaviour also increased gradually. In order to analyze the impact of spatial attention model number on behaviour recognition performance, a comparative experiment is carried out on the first split of HMDB51, and the results are shown in figure 7. As can be seen from figure 7, when the number of spatial attention models is less than 4, the recognition accuracy gradually improves with the increase of the spatial attention model number. When the number of spatial attention models is 4, the performance of behaviour recognition is the best. When the number of spatial attention models is 5, the recognition rate decreases. Due to the limited computer video memory, the experiment cannot run when the number of spatial attention models is greater than 5. Therefore, this paper adopts four spatial attention models to carry out subsequent experiments.

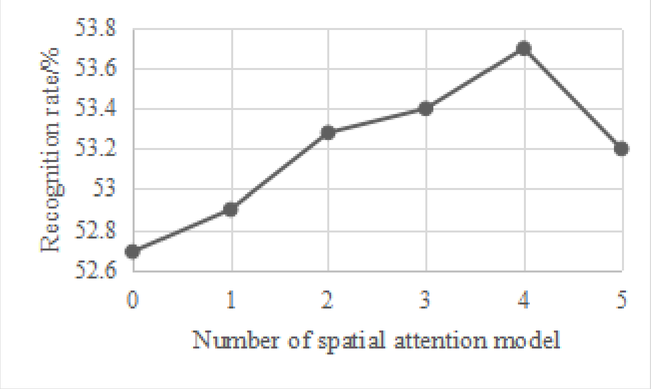


Figure 7. Comparison of recognition accuracy with number of different video segment spatial attention models

Performance analysis of behaviour recognition with different fusion weights

The influence of different fusion weights of appearance flow and motion flow on behaviour recognition performance is analyzed through experiments, and the results are shown in table 1. As can be seen from table 1, the recognition accuracy of single motion flow is higher than that of appearance flow. Fused flow is better than single flow. When the appearance flow and motion flow are combined with 1/4 and 3/4 weight, the behaviour recognition results are the best. Therefore, in this paper, the fusion weight of appearance flow and motion flow is selected as 1:3 for subsequent experiments.

Table 1. Comparison of recognition accuracy(AC) with different fusion weights/%

Method	AC
RGB	53.44
Optical flow	65.04
1/3RGB+2/3Optical flow	69.49
1/2RGB+1/2Optical flow	67.64
1/4RGB+3/4Optical flow	71.91

Performance analysis of behaviour recognition after adding channel attention network

In order to verify the effectiveness of channel-attention network, the recognition accuracy of TSN [10] model with SE-net and TSN model on three data sets are compared, and the same experimental parameters are adopted. The comparison results are shown in tables 2, 3. It can be seen that compared with TSN, the recognition accuracy values of TSN model with SE-net on UCF101 and HMDB51 are improved by 0.8% and 1.2%, respectively. It shows that the channel attention network can highlight the distinguishable channel information in the video, enhance the expression ability of features, and improve the performance of behaviour recognition. As can be seen from table 3, eight classroom behaviors can be recognized more accurately with SE-BN-inception.

Table 2. Comparison of recognition accuracy between TSN and TSN+SE-net on UCF101 and HMDB51

Method	backbone	Recognition accuracy	
		UCF101	HMDB51
TSN	BN-inception	94.9	69.5
TSN+SE-net	SE-BN-inception	95.7	70.7

Table 3. Comparison of recognition accuracy between TSN and TSN+SE-net on real classroom behaviour

Behaviour	Testing A(course A)		Testing B(course B)	
	TSN	TSN+SE-net	TSN	TSN+SE-net
concentration	60.8	62.5	61.7	63.4
interaction	89.4	91.2	90.9	92.8
bowing their heads	91.3	92.7	88.5	91.6
playing with phones	93.6	94.5	92.3	95.6
sleeping	96.3	97.8	96.5	98.1
reading	98.7	99.4	96.1	98.9
writing	95.3	97.2	96.7	98.3
mind wandering	89.2	91.4	90.2	93.4

Comparison analysis with state-of-the-art behaviour recognition methods

In order to further verify the proposed method in this paper, we conduct comparison with some classical behaviour recognition methods, and the results are shown in table 4. As can be seen from table 4, compared with the traditional method IDT [25], the proposed method has a higher recognition accuracy, indicating that the proposed spatio-temporal attention model can effectively extract the key spatio-temporal information in the video and improve the effect of behaviour recognition. The end-to-end structure of the proposed method makes the calculation more concise. Compared with the dual-stream model [7] and the temporal segmentation network (TSN) [10], the proposed method improves the recognition accuracy by 3.2% and 0.8% on UCF101 data set, and 6.5% and 2.5% on HMDB51 data set, respectively. It shows that the spatio-temporal attention model can effectively extract more behaviour features on key frames, and the behaviours in the video can be more accurately recognized by these information. Compared with the TDD[26], the deeply trained C3D network [27], the spatio-temporal residual model ST-ResNet [28], the spatio-temporal pyramid model [29], ARTNet [30] and TSM [31], it can be seen that the proposed method has a better recognition effect. The proposed method takes the dual-flow features, and the recalibration of channel features into account which highlights the key channel information. The proposed spatio-temporal attention model fully mines the key spatio-temporal information of video, it obtains the video features with enhanced expression ability, and establishes the comprehensive behaviour description.

Table 4. Comparison of average recognition accuracy with other methods/%

Method	UCF101	HMDB51	Real
IDT	85.9	57.2	60.3
Two-stream fusion	92.5	65.4	67.2
TSN	94.9	69.4	70.5
TDD	91.4	64.3	65.6
C3D	83.4	57.9	61.5
ST-ResNet	93.4	66.4	65.1
ST-pyramid	94.6	68.9	72.7
ARTNet	95.4	72.0	69.8
TSM	94.5	70.7	72.3
Proposed	95.7	71.9	72.4

Comparison analysis of behaviour recognition methods using attention mechanism

In order to verify the validity of the spatio-temporal attention model proposed in this paper, the proposed algorithm without SE-net is compared with other behaviour recognition methods with attention mechanism. The results are shown in table 5. It can be seen from table 5 that the proposed method in this paper has a higher accuracy. Compared with the temporal attention model [32] generated by the RNN method, the accuracy of proposed algorithm without SE-net on the HMDB51 dataset has been improved by 6.3%. This is because temporal attention only extracts the key frames, while the proposed method not only extracts the key frames, but also pays attention to the saliency areas of motion in the spatial dimension, indicating that the combination of temporal and spatial information can effectively improve the recognition accuracy. The recognition effect of the proposed algorithm without SE-net is better than that of

RSTAN [16] and ISTPAN[33], which indicates that with the same backbone, the spatial and temporal attention model proposed in this paper is simple in structure, but it can effectively extract the key spatial and temporal information of the video. Compared with attention cluster [34], Bi-LSTM attention[17] and R-STAN [35], the proposed algorithm without SE-net has better performance. References [34,17,35] all use ResNet as backbone for behaviour recognition, and ResNet network performance is better than BN-Inception. However, this paper uses BN-Inception as backbone and still gets good recognition effect. This shows that the spatio-temporal attention model proposed in this paper can effectively make up for the deficiency of BN-Inception, it can accurately extract the key spatio-temporal information in the video, and improve the accuracy of behaviour recognition. After adding the SE-net, the recognition accuracy of the proposed method in the three data sets is further improved, indicating that the proposed method can improve the performance of behaviour recognition by calibrating the information of feature channels combined with channel attention network.

Table 5. Comparison of average recognition accuracy with attention-based methods/%

Method	backbone	UCF101	HMDB51	Real
Temporal attention	BN-inception	93.4	65.1	66.3
RSTAN	BN-inception	94.7	70.6	69.5
ISTPAN	BN-inception	94.9	69.7	71.4
Attention cluster	ResNet-152	94.7	69.3	69.7
Bi-LSTM attention	ResNet-152	94.9	72.0	71.9
R-STAN	ResNet-152	94.6	68.8	67.9
Proposed without SE-net	BN-inception	95.4	71.4	70.8
Proposed	SE-BN-inception	95.8	72.0	72.8

6. Conclusion

Traditional behaviour recognition methods ignore the difference of channel information, cannot distinguish video redundant frames, background, etc, which results in the poor feature expression ability and the low recognition rate. In order to improve the efficiency of students in class, this paper proposes a new students behaviour recognition method based on spatio-temporal attention fusion model. In this paper, channel attention is first integrated into the spatio-temporal structure, and channel information is calibrated through the modeling of channel features to improve the ability of feature expression in videos. The temporal attention model and multi-spatial attention model based on CNN are presented to focus on multiple saliency areas of behaviour on the frames to further enhance the feature representation of the video. In this paper, comparison experiments are carried out on UCF101, HMDB51 data sets and real classroom behaviours. Compared with the advanced methods, the proposed method has achieved a higher recognition accuracy. In the future, we will apply more advanced deep learning methods for students behaviour recognition.

Acknowledgements.

The authors appreciate the anonymous reviewers for their meaningful comments.

References

[1] Zhang, Z., Li, W., Zhang, Y. (2021) Automatic Construction and Extraction of Sports Moment Feature Variables Using Artificial Intelligence. *Complexity* 2021(2): 1-13.

[2] A, J., and Yin, S. (2021) A New Feature Fusion Network for Student Behavior Recognition in Education. *Journal of Applied Science and Engineering* 24(2): 133-140.

[3] Stanislav, S., Laura, S., E, Onaindia. (2020) Behaviour recognition of planning agents using Behaviour Trees. *Procedia Computer Science* 176:878-887.

[4] Tong, M., Li, M., Bai, H., et al. (2020) DKD–DAD: a novel framework with discriminative kinematic descriptor and deep attention-pooled descriptor for action recognition. *Neural Computing and Applications* 32: 5285-5302.

[5] Ou, H., Sun, J. (2021) Multi-scale spatiotemporal information deep fusion network with temporal pyramid mechanism for video action recognition. *Journal of Intelligent and Fuzzy Systems* 5:1-13.

[6] Zheng, D., Li, H., and Yin, S. (2020) Action Recognition Based on the Modified Two-stream CNN. *International Journal of Mathematical Sciences and Computing (IJMSC)* 6(6):15-23.

[7] Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016) Convolutional Two-Stream Network Fusion for Video Action Recognition, *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27-30 June 2016, 1933-1941.

[8] Shi, X., Li, Y., Liu, F., et al. (2021) T-STAM: end-to-end action recognition model based on two-stream network with spatio-temporal attention mechanism. *Application Research of Computers* 38(4): 1235-1239.

[9] Dai, W., Chen, Y., Huang, C., et al. (2019) Two-Stream Convolution Neural Network with Video-stream for Action Recognition, *In 2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 14-19 July 2019, 1-8.

[10] Wang, L., et al. (2016) Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In:Computer Vision-ECCV 2016. *Lecture Notes in Computer Science*, vol 9912: 20-36, Springer, Cham.

[11] Zhao, R., Ali, H., and Van der Smagt P. (2017) Two-stream RNN/CNN for action recognition in 3D videos, In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24-28 Sept. 2017, 4260-4267.

[12] Tua, Z., Xie, W., Qin, Q., et al. (2018) Multi-stream CNN: Learning representations based on human-related regions for action recognition. *Pattern Recognition* 79:32-43.

[13] Zhang, J., et al. (2019) Attention-Based Convolutional and Recurrent Neural Networks for Driving Behavior Recognition Using Smartphone Sensor Data. *IEEE Access* 7: 148031-148046.

[14] Hu, J., Shen, L., Albanie, S., et al. (2020) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(8): 2011-2023.

[15] Sharma, S., Kiros, R., Salakhutdinov, R. (2015) Action Recognition using Visual Attention. Neural Information Processing Systems (NIPS) Time Series Workshop, 11 Dec. 2015.

[16] Du, W., Wang Y., and Qiao, Y. (2018) Recurrent Spatial-Temporal Attention Network for Action Recognition in Videos. *IEEE Transactions on Image Processing* 27(3): 1347-1360.

[17] Yang. H., Zhang. J., Li. S., et al. (2018) Bi-direction hierarchical LSTM with spatial-temporal attention for action recognition. *Journal of Intelligent and Fuzzy Systems* 36(1):1-12.

[18] Karim, S., Zhang, Y., Yin, S. et al. (2019) Impact of compressed and down-scaled training images on vehicle detection in remote sensing imagery. *Multimedia Tools and Applications* 78: 32565-32583.

[19] Wang, X., Yin, S., Li, H. (2020) A Network Intrusion Detection Method Based on Deep Multi-scale Convolutional Neural Network. *International Journal of Wireless Information Networks* 27(4): 503-517.

[20] Li, M. (2013) Specifying Usage Control Model With Object Constraint Language. *ICST Transactions on Scalable Information Systems* 13(1).

[21] Junejo, A., Shen, Y., Laghari, A., et al. (2019) Molecular Diagnostic and Using Deep Learning Techniques for Predict Functional Recovery of Patients Treated of Cardiovascular Disease. *IEEE Access* 7: 120315-120325.

[22] Wang, P., Li, W., Gao, Z., et al. (2016) Action Recognition From Depth Maps Using Deep Convolutional Neural Networks. *IEEE Transactions on Human-Machine Systems* 46(4): 498-509.

[23] Hou, Y., Wang, S., Wang, P., et al. (2018) Spatially and Temporally Structured Global to Local Aggregation of Dynamic Depth Information for Action Recognition. *IEEE Access*, 6: 2206-2219.

[24] Wang, H., and Schmid, C. (2013) Action Recognition with Improved Trajectories. *2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, 1-8 Dec. 2013, 3551-3558.

[25] Wang, L., Qiao Y., and Tang, X. (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 7-12 June 2015, 4305-4314.

[26] Li, J., Liu, X., Xiao, J., et al. (2019) Dynamic Spatio-Temporal Feature Learning via Graph Convolution in 3D Convolutional Networks. *2019 International Conference on Data Mining Workshops (ICDMW)*, Beijing, China, 8-11 Nov. 2019, 646-652.

[27] Chen, J., Kong, J., Sun, H., et al. (2020) Spatiotemporal Interaction Residual Networks with Pseudo3D for Video Action Recognition. *Sensors* 20(11):3126.

[28] Wang, Y., Long, M., Wang J., et al. (2017) Spatiotemporal Pyramid Network for Video Action Recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21-26 July 2017, 2097-2106.

[29] Wang, L., Li, W., et al. (2018) Appearance-and-Relation Networks for Video Classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18-23 June 2018, 1430-1439.

[30] Lin, J., Gan C., and Han, S. (2019) TSM: Temporal Shift Module for Efficient Video Understanding. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 27 Oct.-2 Nov. 2019, 7082-7092.

[31] Liu, Z., Tian, Y., and Wang, Z. (2017) Improving human action recognitionby temporal attention. *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, 17-20 Sept. 2017, 870-874.

[32] Du, Y., Yuan, C., Li, B., Zhao L., Li Y., Hu W. (2018) Interaction-Aware Spatio-Temporal Pyramid Attention Networks for Action Classification. *In: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds) Computer Vision – ECCV 2018. ECCV 2018*. Lecture Notes in Computer Science, vol 11220. pp. 388-404. Springer, Cham.

[33] Long, X., et al. (2020) Purely Attention Based Local Feature Integration for Video Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (Early Access) doi: 10.1109/TPAMI.2020.3029554.

[34] Meng, J., Pan, P., Yang, Z., et al. (2020) Degradable and highly sensitive CB-based pressure sensor with applications for speech recognition and human motion monitoring. *Journal of Materials Science*, 55(7): 10084-10094.

[35] Liu, Q., Che, X., and Bie, M. (2019) R-STAN: Residual Spatial-Temporal Attention Network for Action Recognition. *IEEE Access*, 7: 82246-82255.