Journal of Intelligent Systems and Internet of Things
Vol. 18, No. 02, PP. 434-446, 2026
DOI: https://doi.org/10.54216/JISIoT.180230

Application of Real-Time Behavior Tracking Algorithm Combined with Yolov8 in Student Behavior Detection

Xin Bai[1], Madhavi Devaraj[1,*], Zhe Zhang[1]
[1]School of Information Technology, Mapua University, Manila 1002, Philippines
Emails: Xin.Bai@gmail.com; madhavidevaraj@gmail.com; Zhe.Zhang@gmail.com

Abstract

In the intelligent teaching environment, it is indirect and difficult for teachers to capture learners' learning attitudes and behaviors through digital learning behavior data provided by intelligent platforms. The purpose of this paper is to improve the precision of student behavior detection in teaching, and to provide teachers with a more reliable basis for making teaching plans. The Yolov8 algorithm is applied to student behavior recognition, and a bounding box loss function based on dynamic focusing mechanism is introduced to make a balance between samples with good regression quality and poor regression quality. Through experimental analysis, we can see that the real-time behavior tracking algorithm combined with Yolov8 proposed in this paper has a good application effect in student behavior detection. Moreover, it not only improves the precision of student behavior recognition, but also improves the stability of the algorithm, which is conducive to the effective development of subsequent smart teaching models.

Received: March 27, 2025 Revised: June 28, 2025 Accepted: August 30, 2025

Keywords: Yolov8; Real-time detection; Behavioral tracking; Student behavior

1. Introduction

With the development of information technology, video has become an important tool for recording classroom teaching, which can completely reproduce the real classroom situation. The analysis of classroom videos can help teachers find their own advantages and disadvantages in teaching, reflect and revise the teaching process, and thus promote the professional development of teachers. In addition, using certain methods to quantify the classroom situation is helpful to reflect the classroom teaching activities intuitively and evaluate the teaching effect objectively.

Traditional classroom behaviour recognition is mainly carried out manually, which is time-consuming and inefficient. However, with the development of computer vision technology, deep learning networks provide an efficient solution for classroom behaviour recognition by virtue of their advantages of effectively extracting complex features in classroom videos and automatically recognizing student behaviour [1]. At present, the recognition of students' classroom behaviour based on deep learning is in the early stage of research. The main research method is to recognize students' classroom behaviour through facial expressions, human skeleton, head and posture estimation. Although this method can identify students' behaviour, the disadvantage is that it does not pay enough attention to the interaction between students and the objects around them. In a real classroom environment, many classroom behaviours also include the interaction between students and surrounding objects, such as reading, writing, playing with mobile phones, etc. Therefore, analysing the interaction relationship between students and surrounding objects can provide more useful information for classroom behaviour recognition, and improve the precision of classroom behaviour recognition [2].

With the help of computer technology and modern information communication technology to construct online teaching environment, learners' self-regulated learning behaviour in this environment with network as a medium refers to online learning behaviour. Moreover, teachers often use observation, test, interview and other means to analyse the learning process of learners in the traditional teaching environment, and find out that learners with negative learning attitudes, give timely intervention and correct their learning behavior [3].

Learners' online learning behavior is divided and studied by using indicators such as the rumination ratio of video learning data in the intelligent platform, chapter test scores, and topic discussion times, so as to provide a method reference for the process tracking, supervision and evaluation of online courses. This paper aims to improve the precision of student behaviours detection in teaching through intelligent algorithms. The Yolov8 algorithm is applied to student behaviours recognition, and a bounding box loss function based on dynamic focusing mechanism is introduced. Through this research, while improving the precision of student behaviour recognition, it also improves the stability of the algorithm, which is conducive to the effective development of subsequent intelligent teaching models.

2. Related Work

Currently, using machine learning technology to analyse classroom surveillance videos to identify student classroom learning behaviour has become a new research hotspot and has achieved certain research results. For example, reference [4] processed classroom surveillance videos based on the ESRGAN detection network and used YOLOv5s to identify behaviours such as students playing with their phones, attending classes, and sleeping in classroom surveillance images; Reference [5] proposed an improved lightweight network based on MobileNetV2, which uses C-inverse residual blocks instead of traditional modules to improve the recognition accuracy of the network and recognize classroom behaviours such as students sleeping and writing; Reference [6] enhanced the feature extraction ability of the YOLOv5s network by correcting the BN layer, effectively identifying student behaviours such as writing, eating, and listening to classes; Reference [7] identifies typical classroom behaviours of students, such as playing with mobile phones and standing up, based on their skeletal information characteristics. These deep learning based student behaviour research methods usually directly extract feature information of students from images for behaviour classification. The recognition effect of typical classroom behaviours in specific experimental environments is relatively ideal, but actual classroom scenes are generally affected by some objective factors, including different distances of students, similar actions, and occlusion. These objective factors are the difficulties and challenges of behaviour recognition in general classroom monitoring video scenes. Numerous studies have shown that identifying a person's actions and behaviours requires not only detecting a single target object, but also identifying its interaction activities with surrounding objects.

Some researchers are dedicated to the research of visual relationship detection and have made significant progress. They have found that compared to traditional machine vision tasks such as object detection, image segmentation, and action recognition, visual relationship detection focuses more on the semantic relationships between object pairs. Reference [8] proposed a large-scale dataset HICO for human interaction, which has made significant progress in HOI detection technology. HOI detection is mainly divided into two technical routes: single-stage and two-stage: ① The single-stage HOI detection directly detects the interaction behaviour in the image, but its accuracy is low for multi-target recognition. ② The two-stage HOI detection is mainly divided into two research directions: based on multi stream branches and based on graph convolutional neural networks. The detection method based on multi stream branches combines feature extraction, spatial relationships, and other branch networks to form multi stream branches. Reference [9] proposed the BAR-CNN network, which utilizes chain rule decomposition probability network to encode the spatial positional relationship between people and objects; The IPNet network proposed in reference [10] is used to predict the interaction points between humans and objects, and to locate and classify interaction relationships.

The above research methods mainly infer interaction relationships by extracting appearance features and spatial relationships between people and objects, but lack attention to contextual features, and there is still great potential for improving recognition accuracy. Reference [11] proposed the DCANet network, which integrates global contextual features into character interaction detection, improving the accuracy of network detection. The detection method based on graph convolutional neural networks provides a new approach for character interaction detection. This method applies graph neural networks to HOI detection, constructs an analytical graph of the interaction relationships between characters, and uses graph neural networks to capture more contextual features. Reference [12] proposes the DRG network, which utilizes abstract spatial semantics to describe each group of people and objects, and aggregates contextual information in the scene through a dual relationship graph. The Visual Spatial Graph Network (VSGNet) proposed in reference [13] effectively characterizes spatial relationship features by constructing interaction graphs between humans and objects, significantly improving the accuracy of human object interaction recognition. Considering the advantages of two-stage VSGNet networks in terms of recognition accuracy and representation of human object spatial relationships, reference [14] proposes a student classroom behaviour recognition network based on character interaction based on VSGNet, optimizing and improving the target detection module, character interaction relationship construction, and other aspects to achieve classroom behaviour recognition.

The physical environment is the foundation for building smart classrooms, and research focuses on aspects such as classroom layout, lighting, and air quality. For example, reference [15] suggests that lighting settings have a significant impact on student performance, proposes 10 lighting modes for smart classrooms, and designs an

intelligent lighting system based on dynamic scene switching; Reference [16] uses sensors to retrieve features in smart classrooms, and experimental studies have shown that CO2 levels, temperature, humidity, and noise levels are the main environmental factors affecting teaching quality; Reference [17] designed a software agent for identifying and managing smart classroom environments. The agent identifies users in the classroom and records their attendance, controls environmental variables based on conditions such as brightness, noise, and temperature, and takes into account the types of ongoing teaching activities. By comprehensively adjusting these variables, the classroom provides students with a suitable and enjoyable environment; Reference [18] proposed a synchronous online learning system called "Open Smart Classroom", which utilizes web service technology to provide network software control, file uploading, and adding new online classrooms in synchronous live courses; Reference [19] developed a smart classroom interaction system, where students use the Kahoot smartphone loading tool to provide feedback and communication.

The research on the construction framework of smart classroom emphasizes the combination of theory and practice: the smart classroom framework proposed in reference [20] includes elements such as display, management, access, real-time interaction, and tracking; The intelligent classroom architecture proposed in reference [21] consists of three parts, including a scenario aware intelligent classroom prototype, a technology integration model, and supporting measures for the operation of the intelligent classroom; The system applies RFID technology to achieve student attendance management and real-time interaction functions. The research results show that the interaction function has a positive impact on students' learning attitudes, and point out that interactive whiteboard technology is an important feature of smart classrooms [22]; Reference [23] explores the potential of using the Internet of Things to build a smart classroom, which can provide real-time dynamic feedback on course quality, reflecting students' interest in teaching and teachers. This real-time feedback can enable teachers to adjust teaching content during the teaching process to achieve optimal teaching results. The construction of smart classrooms focuses on problem oriented intelligent solutions: Reference [24] proposes a lightweight active spatial service model for smart classrooms, which adopts new technologies such as composite spatial indexing, refresh pause notification service processing methods, and intelligent active update strategies. It can actively provide spatial event services and solve the performance problems of location aware applications in smart classrooms; Reference [25] defines an intelligent classroom based on the multi-agent paradigm, which defines six levels of middleware in the intelligent classroom and proposes two types of proxies, one for describing software components and the other for defining hardware components.

3. Real-time behaviour tracking algorithm combined with Yolov8

3.1 Yolov8 algorithm

The structure of YOLOv8 is shown in Figure 1.

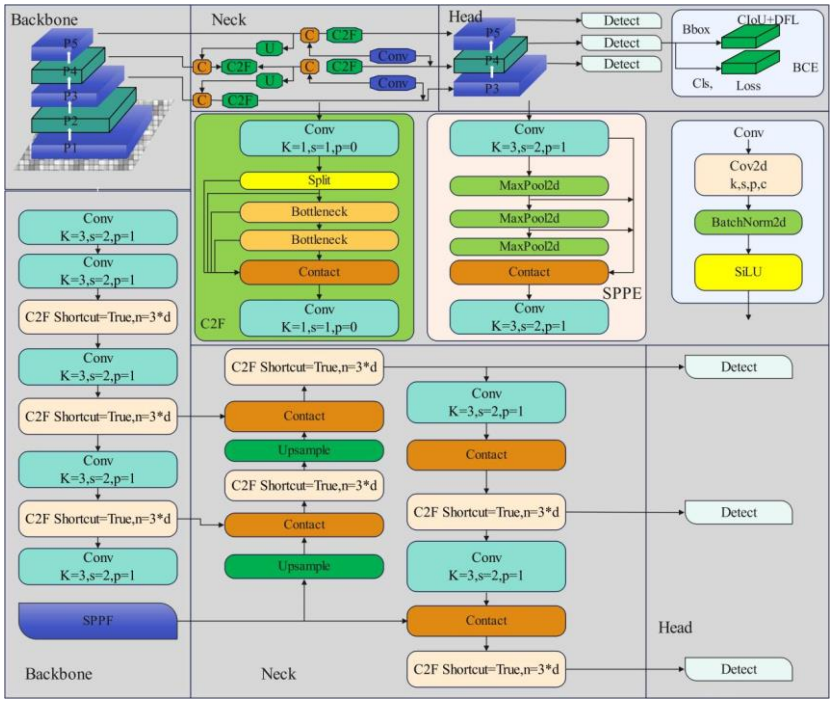


Figure 1. YOLOv8 Framework

The improvement of YOLOv8 in the backbone network part mainly refers to the design idea of the ELAN module in YOLOv7, and replaces the C3 module in YOLOv5 with the C2F module. The C3 module mainly relies on the idea of CSPNet extraction and shunting, and its structure diagram is shown in Figure 2. The gradient flow branch in the C3 module is not fixed. For example, the RepVGG module is used in YOLOv6, and the RepResNet module is used in PP-YOLOE.

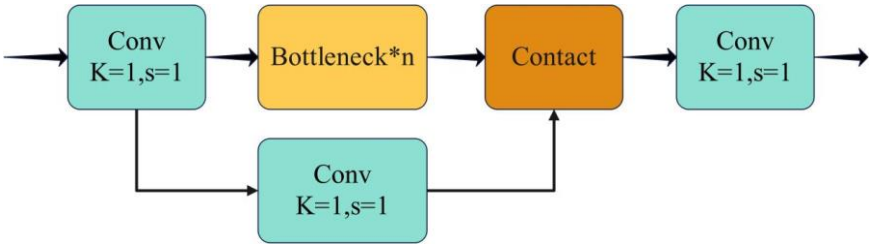


Figure 2. Structure diagram of C3 module

As shown in Figure 3. The main difference from the C3 module in YOLOv5 is that more gradient flow branches are parallel, which obtains richer gradient flow information while ensuring lightweight, and then obtains higher precision and more reasonable delay.

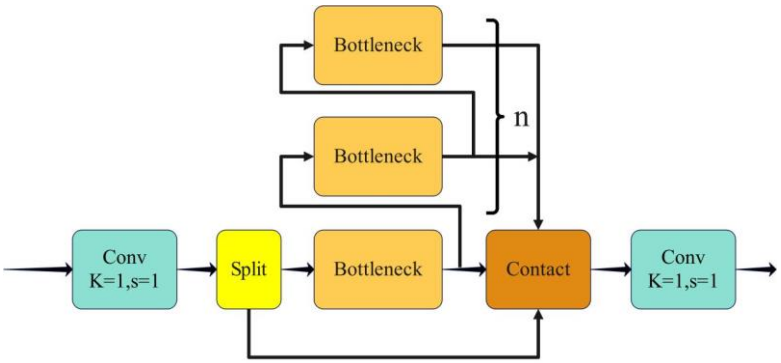
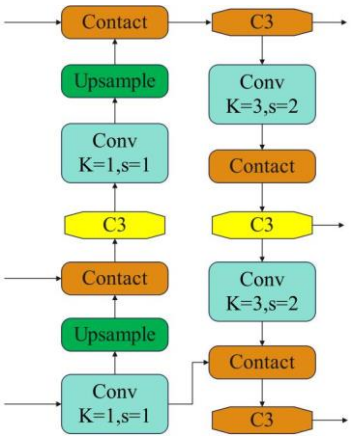
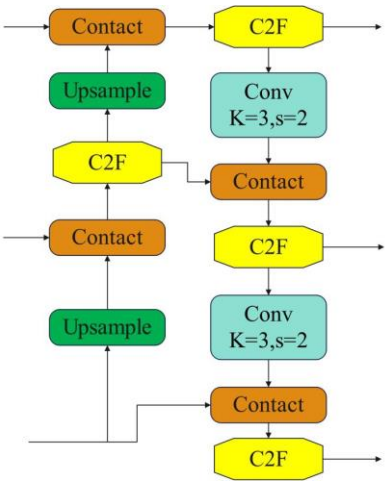


Figure 3. Structure diagram of C2F module

The structure diagram of the Neck part of YOLOv5 and YOLOv8 is shown in Figure 4. YOLOv8 not only replaces the C3 module with the C2F module in the Neck part, but also removes the 1 * 1 convolution connection layer before the two Upsamples, and directly performs the Upsample operation on the features output by the Backbone at different stages.



(a) Structural diagram of the Neck part of YOLOv5



(b) Structural diagram of the Neck part of YOLOv8

Figure 4. Structural diagrams of Neck portions of YOLOv5 and YOLOv8

In YOLOv8, the previous idea of AnchorBased (anchor box based) is abandoned and the AnchorFree (anchor free) method is used. Anchor refers to the approximate position of a pre-set target, which is constrained by setting prior boxes of different sizes and aspect ratios to better match the target object. The AnchorBased method first generates a large number of candidate boxes with different proportions covering almost all positions and scales, then regresses the position of the target relative to the Anchor, and finally corrects it with the corresponding Anchor and regression offset to obtain a more accurate target position. Its advantage is that it improves the accuracy and recall of the model under the constraint of anchor boxes. The disadvantage is that each generated anchor needs to undergo IoU calculation, which increases computational complexity. Moreover, the preset size and proportion of anchors are not flexible enough when detecting objects with significant differences, and need to be changed according to the different datasets. The AnchorFree method in YOLOv8 does not use preset anchors to complete object detection, but converts object detection into keypoint detection. It does not require clustering parameters such as aspect ratio and number of anchors on the current training dataset before training, resulting in stronger generalization ability and a more concise network framework.

3.2 Improvement of target recognition algorithm based on YOLOv8

The neural network can assign different weights according to the relevance and influence of each part of the input information. The attention mechanism can be abstracted into the form of Formula 1, $g(x)$ is the feature processing process, $f(g(x),x)$ is the attention processing process.

$$Attention = f(g(x),x) \tag{1}$$

The soft attention mechanism is a deterministic attention mechanism that does not completely ignore or select any input item. Instead, it learns the attention level of each part of the information through neural networks, and then gives different weights based on the attention level, finally obtaining a comprehensive representation. The soft attention mechanism can be understood as a continuous distribution problem within the [0,1] interval, with the advantage that the process is differentiable and weights can be optimized through backpropagation. The hard attention mechanism emphasizes dynamic changes more than the soft attention mechanism, assigning a non-zero or 1 weight to each input item that is, only focusing on the part of the feature information that is considered to need attention. The hard attention mechanism is a random process, which has the advantage of reducing certain computational costs, but it may lose some feature information that should have been noted. The self-attention mechanism is an attention method that calculates weights based on the interaction between input vectors. It can capture the internal correlation between data and features. The calculation process generally involves first mapping the initial feature map into three vector branches: Query, Key, and Value, then using Query and Key to calculate the correlation between every two input vectors, that is, the attention value. Finally, the weight coefficient matrix obtained by normalizing the attention value is weighted and summed with the Value. But because the self-attention mechanism does not introduce other supervisory information, overfitting is prone to occur. In summary, the soft attention mechanism not only preserves the integrity of input information, but also achieves differentiability, making it convenient for neural networks to train and optimize.

The convolution operation of conventional convolutional neural network will do channel fusion, so the relationship between channels is ignored. Channel domain attention mechanism is an attention method that calculates weights

based on different channels of input vectors. The representative model of channel domain attention mechanism is SENet (Squeeze-and-Excitation Networks), and a large number of subsequent channel domain-based attention mechanisms are mostly improved based on SENet. SENet can be divided into three parts: Squeeze, Excitation, and Scale. Its framework is shown in Figure 5, where Fsq is the Squeeze operation, Fex is the Excitation operation, and Fscale is the Scale operation.

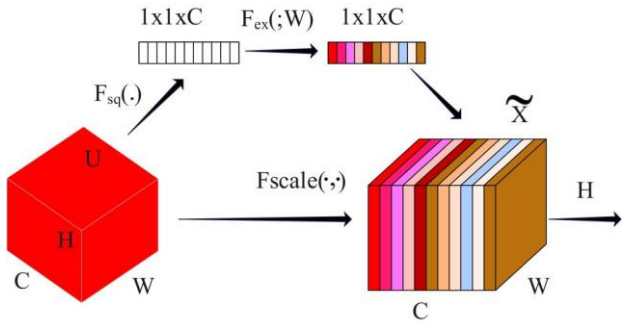


Figure 5. Flowchart of SENet attention module

The Squeeze operation converts a feature map with a size of H * W * C into a vector with a size of 1 * 1 * C through a global average pooling method. The calculation of each value in the vector is shown in formula 2, and Z_c is the values obtained for the corresponding channels. uc is the feature map of each channel, and the size is H * W * 1. The squeeze operation compresses the global information into channel descriptors to mask the spatially distributed information as much as possible, so as to achieve the fusion of global context information.

$$Z_c = F_{sq}(u_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i, j) \tag{2}$$

The first fully connected layer compresses the dimension C to C', and the second fully connected layer is used to restore C' to the original dimension C. The calculation of the gating unit s is as shown in formula 3, where δ is the ReLU activation function, σ is the Sigmoid activation function.According to the experiment, the balance of performance and computational load is achieved when $r = 16$.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(g(W_2 \delta(W_1 z))) \tag{3}$$

The Scale operation assigns the learned weight values of each channel to the input features, and obtains the final feature map. The calculation formula is shown in Formula 4, where $\tilde{x}$ is a feature map of a feature channel in $\tilde{X}$ and S_c is a scalar value in the gating unit s.

$$\tilde{x}_c = F_{scale}(u_c, S_c) = S_c \cdot u_c \tag{4}$$

Most convolutional neural networks use maximum pooling or mean pooling operations to compress data to reduce the amount of computation. This operation directly combines information without distinguishing key information. The representative model of spatial domain attention mechanism is GENet (Gather-Excite Networks). The GENet framework is shown in Figure 6, ξ_G and ξ_E are the Gather and Excite operators defined for GENet.

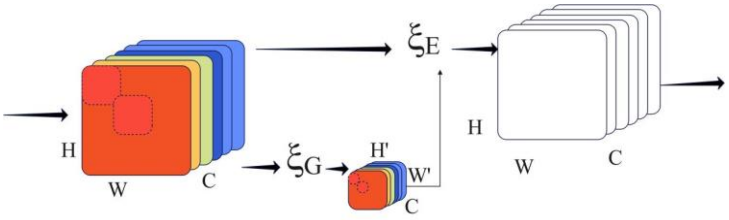


Figure 6. Flowchart of GENET attention module

ξ_G is used to aggregate neuronal responses over a given spatial range, that is, an input feature map of size $H * W * C$ is converted into an output of size H '$* W$' $* C$ by one or more convolution operations. where the relationship of H, W, H ', and W' is shown in formula 5, and e is the selected range ratio.

$$\begin{cases} H' = \dfrac{H}{e} \\ W' = \dfrac{W}{e} \end{cases} \tag{5}$$

For an arbitrary input feature graph x, ξ_G satisfies formula 6, where $l(u,\mathrm{e}) = eu + \delta, \delta \in \left(-\dfrac{2e-1}{2}, \dfrac{2e-1}{2}\right)^2$, $u \in \{1,...,H'\} \times \{1,...,W'\}, c \in \{1,...,W'\}$, and $\odot$ is Hadamard product and $1\{\cdot\}$ is the indication tensor.

$$\xi_G(x)_u^c = \xi_G\left(x \square\ 1_{l(a,c)}^c\right) \tag{6}$$

ξ_E is used to combine aggregated and raw inputs to produce outputs that match the dimensions of the raw inputs, and the calculated formula is shown in Formula 7. Among them, $f: RH' * W' * \mathrm{C} \rightarrow [0,1] H * W * C$ is responsible for rescaling and assigning the mapping of the signals in the aggregation.

$$\xi_E(x, \hat{\mathrm{x}}) = x \square\ f(\hat{x}) \tag{7}$$

In Formula 8, where σ is a sigmoid function, $Favgc$ and $Fmax$ are the characteristics of the average pooling and maximum pooling, and W_0 and W_1 are the weight of MLP.

$$\begin{aligned} M_c(F) &= \sigma\left(MLP\left(AvePool(F)\right) + MLP\left(MaxPool(F)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1\left(W_0\left(F_{mac}^c\right)\right)\right) \end{aligned} \tag{8}$$

The spatial attention calculation is shown in formula 9, where $f^{7\times7}$ is a convolution operation with a convolution kernel size of 7 * 7.

$$\begin{aligned} M_S(F) &= \sigma\left(f^{7\times7}\left(\left[AvePool(F); \mathrm{Max}\ Pool(F)\right]\right)\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F_{avg}^s; F_{\max}^s\right]\right)\right) \end{aligned} \tag{9}$$

3.3 Student behavior testing

The SoftMax loss function is just suitable for students' classroom behavior recognition.In formula (10), i is the number of output nodes, n is the number of categories, and p_i is the probability value of the output node corresponding.

$$p_i = \frac{e_i}{\sum_{k=1}^{n} e_k} \tag{10}$$

The larger the value of p_k, the better the result. The cross-entropy loss is introduced here, which is to calculate the gap between the model value and the real value. The formula is shown in formula (11), k is the behavior label value of the corresponding category, n represents the number of behavior categories, and y represents the marked behavior categories, and y_k is the predicted probability value normalized using the SoftMax function.

$$Loss = -\sum_{k=1}^{n} y_k \log p_c \tag{11}$$

Batch normalization can speed up the convergence speed and improve the precision of the network. Regulates the data input to the network by adjusting and scaling activation.

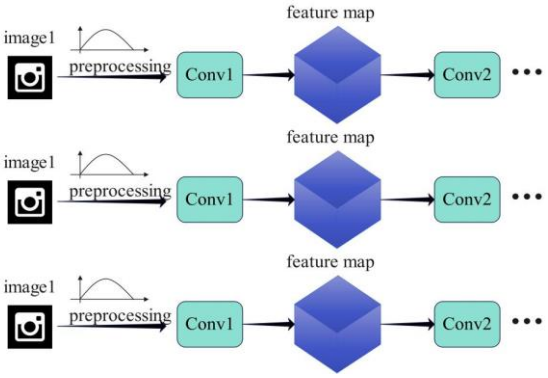


Figure 7. Application of batch normalization

The advantages of batch normalization are: (1) it greatly reduces the contingency caused by artificial parameter selection, abandons the use of Dropout algorithm and cancels the L2 regular term parameter, or replaces it with a smaller L2 regular term constraint parameter; (2) it reduces the trouble of frequently manually adjusting the learning rate; (3) it avoids the use of local normalization operations, and BN itself has the function of normalization; (4) it normalizes the distribution of the original data, which alleviates the overfitting to a certain extent. Therefore, all the image recognition networks in this paper use batch normalization processing.

For each student in the video sequence, the possible motion behavior in the video is found by continuously detecting the gray level change of each pixel in the time window of 8 frames. The gray level change value in these 8 frames is calculated by the accumulation result of the gray level value subtracted from two consecutive frames, so that the coordinate (x, y) of the point with the largest gray level change is obtained. Taking this coordinate as the center, the ROI region containing only each student is cropped, and the size is just 100 x 100. We call this ROI region as ST-ROI. The 8 ST-ROIs are converted into row vectors and stacked column by column, and the stacking formula is shown in formula (12).

$$ST = \begin{bmatrix} t_{11} & \cdots & t_{1p} \\ \cdots & \cdots & \cdots \\ t_{w1} & \cdots & t_{wp} \end{bmatrix} \tag{12}$$

A fast Fourier transform (FFT) is performed on the matrix ST in the column direction, and ft_i represents the i-th row of the matrix ST after the fast Fourier transform. Finally, a transformed matrix F can be obtained, as in formula (13). Through this method, we can extract the pixel position of the 8-frame image which has a large change in gray level.

$$F = ABS\left(FFT\left(ST\right)\right) = ABS\begin{bmatrix} ft_i \\ \cdots \\ ft_w \end{bmatrix} \tag{13}$$

The transformed matrix F is saved with a red channel, the second row of frame information in the selected matrix ST is saved with a blue channel, and the ST-ROI is saved with a green channel. Thus, a spatio-temporal feature information image based on 8 consecutive RGB images is generated. The continuous RGB frame images of each behavior are processed by the above method, and the following spatio-temporal feature information images can be obtained, as shown in Figure 8.

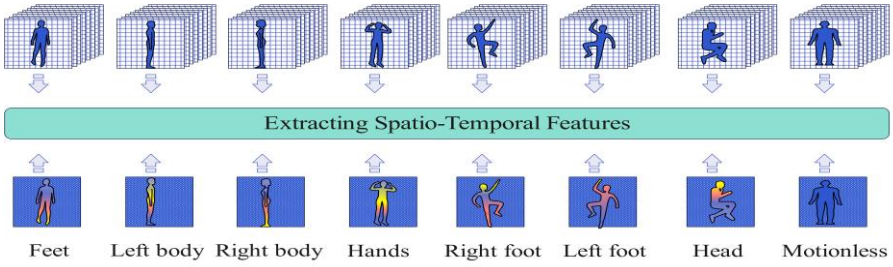


Figure 8. Image of spatio-temporal characteristics of student behavior

4 Experimental analysis

4.1 Experimental results

The confusion matrix is usually used and the loss value, precision and F1 score are calculated. Therefore, in the performance evaluation of the behavior recognition model, we use the confusion matrix, loss value and F1 score as evaluation indicators. F1 score: F1 score takes into account precision and recall, and is the weighted average of the two, and its value range is [0, 1].

During the experiment, we train the models of five convolutional neural networks, which are: (1) AlexNet network; (2) VGG16 network; (3) GoogLeNet network; (4) YOLOv5 algorithm network; (5) YOLOv8 algorithm network. All networks are trained from scratch to facilitate ablation experiments and comparison experiments. Next, the performance of each model is evaluated according to indicators such as the precision, loss value, F1 score and confusion matrix of the training set and verification set. Table 1 is the F1 score for individual network models.

Table 1: F1 scores for individual network models

	AlexNet	VGG16	GoogLeNet	YOLOv5	YOLOv8
Daze	89.83%	88.83%	87.57%	95.19%	97.25%
Sleep;	85.63%	83.34%	81.96%	89.31%	91.83%
Whisper	90.64%	89.50%	88.09%	94.38%	98.13%
Walk	88.61%	85.85%	83.91%	93.83%	92.41%
Cooperation and Exchange	87.46%	85.23%	80.68%	94.51%	95.71%
Answer questions;	93.87%	86.72%	84.57%	90.84%	96.22%
Take notes	81.87%	80.10%	78.86%	93.12%	95.74%
Violation of rules and regulations	89.04%	87.61%	84.46%	92.07%	96.61%
Average F1 score	88.37%	85.90%	83.76%	92.90%	95.49%

At the same time, this paper also counts the weights of each network model, which is also convenient for deploying the network model to mobile devices or hardware devices in the future, and provides a comparison choice for subsequent actual device applications. The weights of each network model are shown in Table 2.

Table 2: Weight of each network model

Model	Weight (MB)
AlexNet	55.143
VGG16	265.5477
GoogLeNet	39.1149
YOLOv5	97.119
YOLOv8	99.297

In order to verify the practical applicability of this study, we select 4 new evaluation videos, the number of students in each video is 30, 40, 60 and 100, and the duration of each video is 60 minutes. The method proposed in this paper is used to automatically record the number of behaviors of each student in this period, the precision of model records, the precision of manual records, the deviation of model records, the average precision and standard deviation of model records. The correct rate is tested by observing the behavior interval with the naked eye by

multiple professionals, that is, the detection of the behavior interval is considered correct only when it meets the manual statistical standard. Moreover, all test results are shown in Table 3.

Table 3: Identification results of student behavior

Video number	Student Number	Behavior number	Precision of model record	Precision of manual recording	Deviation from manual recording
1	1	480	96.65%	93.53%	3.13%
	2	564	97.23%	94.78%	2.45%
	3	610	96.91%	97.69%	-0.78%
2	1	336	95.70%	96.51%	-0.80%
	2	555	95.17%	93.84%	1.33%
	3	685	98.17%	95.76%	2.41%
	4	343	96.35%	97.75%	-1.41%
3	1	281	92.34%	97.15%	-4.81%
	2	408	97.45%	95.43%	2.02%
	3	425	95.06%	94.02%	1.04%
	4	225	90.76%	94.24%	-3.47%
	5	127	92.36%	94.77%	-2.42%
	6	341	95.30%	93.59%	1.70%
4	1	432	96.42%	94.01%	2.41%
	2	353	94.76%	95.52%	-0.75%
	3	404	97.40%	94.36%	3.04%
	4	186	91.53%	96.21%	-4.68%
	5	339	97.35%	96.11%	1.24%
	6	287	94.75%	94.94%	-0.19%
	7	505	96.26%	95.93%	0.33%
	8	162	89.03%	96.56%	-7.53%
	9	359	95.33%	94.76%	0.56%
	10	543	96.40%	94.93%	1.47%
Average Precision	95.16%				
Standard deviation	2.81%				

For the above four student behavior videos used for evaluation, this paper also compares the time consumption of the manual frame-by-frame recording statistical method and the automatic statistical method using the behavior recognition model in this paper. The results are shown in Table 4.

Table 4: Statistical time-consuming comparison between manual recording and model recognition

Video number	Video duration (h)	Total number of acts	Manual recording time (h)	Model recognition statistic duration (h)
1	1.5	1654	16.05	1.43
2	1.5	1919	17.33	1.56
3	1.5	1806	16.26	1.50
4	1.5	3570	31.83	1.61

4.2 Analysis and discussion

As shown in Table 1, the F1 scores of each network model on different behavior categories are shown. From the data in the table, it can be seen that the deeper network models YOLOv5 and YOLOv8 still have higher F1 scores. The reason is that these two networks are better for image feature extraction and learning. Although the previous shallow network models such as AlexNet, VGG16 and GoogLeNet can also achieve a good F1 score, they are not sufficient for image feature learning. Especially, GoogLeNet has the worst performance effect. It may be because the multi-branch structure in its network structure does speed up the training speed, but the feature information learning is not rich enough, so the recognition effect is not good. The behavior appears more frequently, and the number of samples in the training set is more, so the recognition effect is more accurate.

From the data in Table 3, it can be seen that the average precision rate of behavior recognition of all students is 95.16%, and the standard deviation is 2.81%, which meets the requirements of automatic behavior recognition recording and quantification, indicating the practicability of the method in this paper.

As shown in Table 4, it can be seen that for videos of the same duration, the more the total number of behaviors, the longer the time consumed by manual statistics, and when watching videos for a long time, people’s eyes will tire and cause errors. When using the statistical method of the computer model, we only need to manually select the video or folder that needs to be processed to automatically identify and count the student behavior in the video. Therefore, this will greatly save time and is not easy to make mistakes. This is also the role and value of computer vision technology applications in improving students’ learning efficiency.

Through the above analysis, we can see that the real-time behavior tracking algorithm combined with Yolov8 proposed in this paper has a good application effect in student behavior detection, and it not only improves the precision of student behavior recognition, but also improves the stability of the algorithm.

The model in this article can be combined with specialized databases to recognize various behaviors of students, rather than just recognizing a certain scenario. By constructing a video dataset of student practical courses in practical learning scenarios and researching and improving related spatiotemporal action localization algorithms, student news from different scenarios can be obtained for the database. By combining with the campus intelligent video monitoring system, students' behavior can be intelligently recognized

Video data provides temporal information, which requires algorithms to capture dynamic features such as the start, progress, and end of actions. Therefore, video action recognition not only needs to analyze spatial features, but also needs to handle changes in time series, which increases the complexity of the task. When processing video data and detecting and classifying the behavior of characters in videos, the network needs to consider the relationship between the previous and subsequent frames, that is, the temporal features. If only one frame of the image is considered, it can easily lead to ambiguity in actions, such as in the dynamic process of sitting to standing and standing to sitting, it is difficult to determine whether the character's behavior is sitting or standing if only the middle frame is viewed. Therefore, being able to extract temporal features is crucial for behavior recognition and detection tasks based on video data. From this, it can be seen that the model in this article can be applied to all behavior recognition of campus students, as well as to action recognition in other fields

Use transfer learning to enhance the generalization of the model and introduce appropriate loss functions to improve the accuracy of the detection network. In order to further enhance the detection ability of the model in school management for small and overlapping goals of students, as well as real-time detection of various behaviors

of students, the extension effect of the model can be effectively improved on the basis of transfer learning enhancement model. It can also be combined with the smart campus platform of universities through embedded methods to effectively improve the management effect of smart campus.

The model in this article will monitor students' real-time behavior, which inevitably leads to monitoring of their daily life and other behaviors. Therefore, in order to solve ethical and privacy issues, the model in this article only connects to the campus LAN and stores data through the solid-state drive inside the campus. It only temporarily saves behaviors that meet the characteristics of the database, making it convenient for teach management personnel to proceed with the next step of processing And only personnel with administrator privileges can process video data in this system, and other personnel cannot enter the system to view any data without authorization

4. Conclusion

In the full information recognition method of student behavior, it is mainly divided into two categories. The first category is to track and locate the key points of students' limbs first, and then design mathematical algorithms based on the relative position information of the key points to estimate students' behavior, which can be expressed as student posture estimation. The second is to extract the spatio-temporal characteristic information of students' learning behavior and generate spatio-temporal characteristic information images. According to the spatio-temporal characteristic information images of behaviors, various types of combing behaviors can be clearly characterized. Finally, a convolutional neural network is constructed to identify each combing behavior. It is obvious that the latter method has a higher precision rate of identifying behaviors. Therefore, this paper uses YOLOv8 algorithm and traditional image processing technology to extract and generate spatio-temporal feature information images of each behavior at the same time.

When behavior recognition methods are applied to more types of behavior recognition than just in the classroom, it is necessary to enhance the generalization of the network model. For example, the application of intelligent recognition technology to the daily management of schools is also the follow-up research direction.

References

[1] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu, "Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology," *Journal of Educational Computing Research*, vol. 58, no. 1, pp. 63–86, 2020.

[2] TS and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Education and Information Technologies*, vol. 25, no. 2, pp. 1387–1415, 2020.

[3] G. Gorgun and O. Bulut, "Identifying aberrant responses in intelligent tutoring systems: An application of anomaly detection methods," *Psychological Test and Assessment Modeling*, vol. 64, no. 4, pp. 359–384, 2022.

[4] F. Noorbehbahani, A. Mohammadi, and M. Aminazadeh, "A systematic review of research on cheating in online exams from 2010 to 2021," *Education and Information Technologies*, vol. 27, no. 6, pp. 8413–8460, 2022.

[5] W. Xu and F. Ouyang, "The application of AI technologies in STEM education: A systematic review from 2011 to 2021," *International Journal of STEM Education*, vol. 9, no. 1, pp. 59–70, 2022.

[6] Horvers, N. Tombeng, T. Bosse, A. W. Lazonder, and I. Molenaar, "Detecting emotions through electrodermal activity in learning contexts: A systematic review," *Sensors*, vol. 21, no. 23, pp. 7869–7880, 2021.

[7] H. El Aouifi, M. El Hajji, Y. Es-Saady, and H. Douzi, "Predicting learner's performance through video sequences viewing behavior analysis using educational data-mining," *Education and Information Technologies*, vol. 26, no. 5, pp. 5799–5814, 2021.

[8] M. Awais *et al*., "LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19," *IEEE Internet of Things Journal*, vol. 8, no. 23, pp. 16863–16871, 2020.

[9] R. Baker *et al*., "The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: Opening the black box of learning processes," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1–24, 2020.

[10] S. Raj and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 994–1004, 2020.

[11] J. Chen, M. Abbod, and J. S. Shieh, "Pain and stress detection using wearable sensors and devices—A review," *Sensors*, vol. 21, no. 4, pp. 1030–1040, 2021.

[12] Y. Ding, X. Chen, Q. Fu, and S. Zhong, "A depression recognition method for college students using deep integrated support vector algorithm," *IEEE Access*, vol. 8, pp. 75616–75629, 2020.

[13] A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interactive Learning Environments*, vol. 30, no. 8, pp. 1414–1433, 2022.

[14] S. Sengupta and A. Vaish, "Social networking mood recognition algorithm for conflict detection and management of Indian educational institutions," *Social Network Analysis and Mining*, vol. 10, no. 3, pp. 1–13, 2020.

[15] P. Chikersal *et al*., "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection," *ACM Transactions on Computer-Human Interaction*, vol. 28, no. 1, pp. 1–41, 2021.

[16] Z. Guo *et al*., "Robust spammer detection using collaborative neural network in Internet-of-Things applications," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9549–9558, 2020.

[17] Behera *et al*., "Associating facial expressions and upper-body gestures with learning tasks for enhancing intelligent tutoring systems," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 236–270, 2020.

[18] Z. Hu *et al*., "Statistical techniques for detecting cyberattacks on computer networks based on an analysis of abnormal traffic behavior," *International Journal of Computer Network and Information Security*, vol. 12, no. 6, pp. 1–11, 2020.

[19] Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, pp. 552–566, 2021.

[20] J. C. Paiva, J. P. Leal, and Á. Figueira, "Automated assessment in computer science education: A state-of-the-art review," *ACM Transactions on Computing Education*, vol. 22, no. 3, pp. 1–40, 2022.

[21] R. Harper, T. Bretag, and K. Rundle, "Detecting contract cheating: Examining the role of assessment type," *Higher Education Research & Development*, vol. 40, no. 2, pp. 263–278, 2021.

[22] S. Chai, X. Wang, and C. Xu, "An extended theory of planned behavior for the modelling of Chinese secondary school students' intention to learn artificial intelligence," *Mathematics*, vol. 8, no. 11, pp. 2089–2101, 2020.

[23] Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, pp. e1355, 2020.

[24] Channa *et al*., "The rise of wearable devices during the COVID-19 pandemic: A systematic review," *Sensors*, vol. 21, no. 17, pp. 5787–5799, 2021.

[25] K. Sharma and M. Giannakos, "Multimodal data capabilities for learning: What can multimodal data tell us about learning?" *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1450–1484, 2020.