Research Article

Liu Yang*

Application of online teaching-based classroom behavior capture and analysis system in student management

https://doi.org/10.1515/jisys-2023-0236
received November 07, 2023; accepted November 26, 2023

Abstract: Analyzing online learning behavior helps to understand students' progress, difficulties, and needs during the learning process, making it easier for teachers to provide timely feedback and personalized guidance. However, the classroom behavior (CB) of online teaching is complex and variable, and relying on traditional classroom supervision methods, teachers find it difficult to comprehensively pay attention to the learning behavior of each student. In this regard, a dual stream network was designed to capture and analyze CB by integrating AlphaPose human keypoint detection method and image data method. The experimental results show that when the learning rate of the model parameters is set to 0.001, the accuracy of the model is as high as 92.3%. When the batch size is 8, the accuracy of the model is as high as 90.8%. The accuracy of the fusion model in capturing upright sitting behavior reached 97.3%, but the accuracy in capturing hand raising behavior decreased to only 74.8%. The fusion model performs well in terms of accuracy and recall, with recall rates of 88.3, 86.2, and 85.1% for capturing standing up, raising hands, and sitting upright behaviors, respectively. And the maximum $F1$ value is 0.931. The dual stream network effectively integrates the advantages of two types of data, improves the performance of behavior capture, and improves the robustness of the algorithm. The successful application of the model is beneficial for teachers' classroom observation and research activities, providing a favorable path for their professional development, and thereby improving the overall teaching quality of teachers.

Keywords: online teaching, keypoint detection method, RGB images, double flow network, behavior capture

1 Introduction

With the development of network technology, Internet platforms have accelerated the digital transformation of teaching models. Online learning (OT) platforms provide teachers and students with flexible learning venues, open educational resources, and social learning modes. Compared with traditional offline classroom education, OT is convenient and flexible, with the obvious advantages of personalized learning, and the scale of OT is gradually expanding. OT platforms record and save a large number of real-time classroom video images, including a large number of students' classroom behavior (CB). The analysis of CB can further understand the learning situation of students, so that teachers can give timely feedback and personalized guidance to improve the learning effect and the quality of teaching. However, it is difficult to supervise the learning discipline in online classroom, and it is difficult for teachers to observe and collect students' CBs by their own senses, so it is difficult to analyze the inner mechanism behind students' CBs in real time and cannot form timely feedback. And the CB is complex and variable, the recording of video images is often entrained with

*** Corresponding author: Liu Yang,** Department of Foreign Languages, Hebei University of Architecture, Zhangjiakou 075000, China, e-mail: yl1541@hebiace.edu.cn

more interference background and other noise, which affects the artificial observation and identification of CB, and influences the development of teaching activities [1,2]. Therefore, the regulation of OT behaviors should consider seeking more scientific and effective methods to promote changes in the teaching industry and improve the quality and efficiency of teaching. The development of artificial intelligence, image processing, deep learning, and other fields provides new solutions for the regulation of OT classroom, and teachers can use information technology to detect, process, and analyze CBs [3,4]. Based on the big data statistics of OT platforms, changes in students' CBs obtained from video can be analyzed using artificial intelligence visual technology for further analysis. However, in the face of multi-obscured and complex classroom background environment, the applicability of the existing behavior capture (BC) and analysis model is obviously insufficient, and there is still much room for improvement in the capture accuracy of various behaviors, such as students' verbal communication and body movements.

Human keypoints detection (HKD) is one of the general tasks of computer vision, which can be used as human behavior tracking and gait recognition. In order to standardize the management of student behaviors in online teaching classrooms, the study introduces the HKD method, fuses RGB image data to construct a dual-stream network that can capture and analyze CBs, and this model innovatively fuses the behavioral features of the two branches for recognition, and uses a behavioral annotation tool for keypoints. This research is expected to improve the technical defects of the existing BC and analysis algorithms, and make improvement research in the aspects of capture accuracy and adaptation to complex environments. The study is divided into four parts, the first part of which summarizes and concludes the recognition and analysis of CBs and general behaviors at home and abroad; the second part proposes the CB capture and analysis model based on the HKD method constructed by the study; the third part conducts the performance test and validation of the algorithm; and the fourth part summarizes and concludes the results of the study. It is expected that the CB capture and analysis model of this research can accurately analyze OT behaviors and promote the informatization process and reform of online education.

2 Related works

For the purpose of enacting educational changes and fostering the creation of new online education, the identification and study of CB are crucial. Several professionals and academics have conducted study on the recognition and analysis model of behavior. Lu et al. proposed a posture guiding model that integrates heat and color images of the human body and employs keypoint action features for individual image behavior detection in order to investigate an algorithmic model that can be used to detect learning distractions and offer early warnings. The model performs exceptionally well on the dataset [5] and a keypoint gating module was included to weight the discrimination of keypoint characteristics. A solution for this issue was put up by Xiao et al. in the form of an attention-based deep neural network strategy for human behavior recognition. Combining a lightweight attention block with a channel attention block improved the model's representativeness, and tests on two publicly available datasets confirmed the strategy's efficacy, with the method outperforming other approaches and obtaining an accuracy of 98.48% [6]. Shi and colleagues developed an improved alphabet-based underground behavior detection system with Cycle-Generative Adversarial Networks to defog and enhance images in subterranean surveillance in an effort to reduce the frequency of safety mishaps caused by miners' risky behavior. In accordance with the findings, the logistic tent chaos mapping-enhanced whale optimization technique had more substantial convergence than other optimization algorithms, and its identification accuracy was 9.1% greater than that of the unoptimized model [7]. Current smart home systems use a variety of sensors, but human interaction is necessary for collecting and predicting user behavior. A new automatically annotated user behavior prediction model was proposed by Zhang et al. that combines a behavior prediction model based on long and short-term memory networks with a behavior recognition model for discontinuous solution sequential sequence mining. The model's efficacy in recognizing user behavior was experimentally confirmed [8].

To accomplish the task of multi-user behavior recognition, An et al. proposed a divide-and-conquer dynamic memory network model. The scholars used gated cyclic units to solve the consistency problem of different behaviors at the data level and extended the model memory based on the idea of dynamic memory networks. Experimental results showed that the model performs well in two dimensions, accuracy and recall, and can improve recognition accuracy [9]. Point cloud recognition and localization is still challenging for applications in complex scenes, and to address this thorny problem, Wu et al. proposed an efficient and highly compatible correspondence grouping technique implemented through correspondence ranking, clustering, and extension operations. This approach is applicable to multi-target recognition tasks, and validation results on four different datasets demonstrated the efficiency and effectiveness of the technique [10]. Similarly, Zhu et al. proposed a new iterative nearest-point algorithm combined with distributed weights to enhance the reliability and robustness of non-collaborative target localization. The results show that the algorithm can effectively suppress interference points and improve the accuracy of non-cooperative target pose estimation. The average error of the angle is better than 0.88 degrees when the number of interference points reaches about 700 [11]. To improve the accuracy of the Levenberg–Marquardt-based pose estimation algorithm, Bilal et al. developed an eye-hand camera-based pose estimation system. The accuracy of the estimated pose was also improved using long and short-term memory neural networks and sparse regression, and the proposed method was compared with the extended Kalman filter. Experimental results demonstrated that both methods significantly improved the pose estimation accuracy and precision of the vision-based system during robot machining, with absolute position errors of 5.47, 2.9, and 2.05 mm on average for machining [12]. To safeguard the privacy of the user's behavior, Lin developed an algorithmic network for monitoring behavior over a limited area, implemented with a local object tracking technique based on a multi-reflective infrared sensor array. The experiments were conducted with caregivers in a healthcare facility ward and showed a 99% correct recognition rate for 26 activities, of which 4 were caregiving activities, 16 were daily activities of patients in bed, and 6 were getting out of bed [13].

Video image processing is an important research area in the field of computer vision. Sharifi and Amini combined particle swarm algorithm and multivariate correlation vector regression to design an optimized feature algorithm for extracting image estimation from polarimetric synthetic aperture radar images. The experimental results showed that the method has the least error with a root mean square error of 39.17, a mean absolute error of 36.50, and a mean error of 11.59. Continuing on this basis, a multivariate correlation vector regression method was designed to be used as the core module of a Bayesian model for estimating aboveground biomass in the Hyrcanian forests in Iran. Experimental results showed that this method has higher accuracy compared to models such as multivariate linear regression and multilayer perceptron neural networks. After that, Sharifi accomplished the classification and identification of satellite images using correlation vector machine for determining over detected and under detected areas in flooded areas, and the correlation vector machine achieved a classification accuracy of 0.89.This study confirmed the contribution of satellite radar imagery in the detection and delineation of water operations in flooded hazardous areas. Kosari et al. devised a performance-based design strategy for a fast sizing method that minimizes the complexity and length of time of the performance sizing process for satellite state determination and control subsystems.

In conclusion, there is still much room for improvement in the recognition and detection of human behavior, and various optimizations and enhancements are still required with regard to the algorithm's recognition accuracy and robustness in adjusting to the complexity of the environment. Meanwhile, related technologies in the field of image processing provide technical ideas for behavioral capture recognition in video images, and it is relatively uncommon to also consider different methods for recording and analyzing student behavior in the classroom, although doing so has wider ramifications for the creation of such models.

3 Construction of CB analysis algorithms based on data fusion

3.1 Design of CB capture algorithms based on multiple data types

This research employs the HKD approach to locate keypoints and successfully gather accurate human behavioral skeleton data on the one hand, in hopes of creating a more precise and scientific learning BC and analysis system. On the other hand, a dual flow BC and analysis system is created based on the dual flow approach by integrating two sources of data: human keypoints and RGB images.

Considering the large classroom population, the study used the well-established multi-person Alpha Pose HKD algorithm to tag student CBs recorded in online instructional videos and used the Alpha Pose algorithm with the multi-person pose estimation Regional Multi-person Pose Estimation (RMPE) framework. The RMPE framework structure is shown in Figure 1 [14,15]. As shown in Figure 1, the RMPE framework mainly consists of the Symmetric Spatial Transformer Network (SSTN), the Parametric Pose No-Maximum-Suppression (PP-NMS), and the Pose Guided Border Generator Pose Guided Proposals Generator (PGPG). Multi-human target detection is done first, SSTN obtains a single human detection frame, then PP-NMS is used to measure the similarity between the poses, and finally the data are augmented with PGPG, which increases the effective pose training samples by dense sampling and the algorithm has a high accuracy. Keypoints are extracted by Alpha Pose HKD algorithm and then keypoints are connected to label the skeleton behavior. In hopes to prevent difficulties in labeling due to the presence of occlusion or small targets, the You Only Look Once v3 (YOLO v3) target detection method is obtained by modifying the structure of the GoogLeNet network with the inception module. And YOLO v3 is more effective in detecting small targets and improves the convergence speed and gradient disappearance problem of the network.

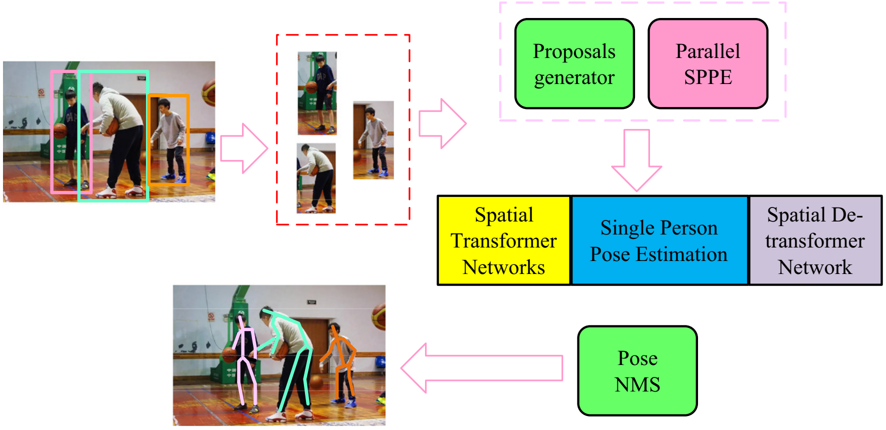
Proposals generator
Parallel SPPE
Spatial Transformer Networks
Single Person Pose Estimation
Spatial De-transformer Network
Pose NMS

Figure 1: RMPE framework diagram.

The values of the horizontal and vertical coordinates of the keypoints and the confidence level of the keypoints were retrieved and subsequently the coordinates of the extracted keypoints were normalized so that the model focused on modeling the relationship between the positions of different keypoints in the human body. The normalization process is described in equation (1), using the way in which the joints are connected as the basis for classifying different capture behaviors.

$$\begin{cases} W = X_{\max} - X_{\min} \\ H = Y_{\max} - Y_{\min} \\ X_n = \dfrac{x_i - X_{\min}}{W} \\ Y_n = \dfrac{y_i - Y_{\min}}{H} \\ C_n = C_i. \end{cases} \tag{1}$$

The keypoints' maximum and minimum horizontal and vertical coordinate values are represented in equation (1) by the letters $X_{\max}$, $X_{\min}$, and $Y_{\max}$, $Y_{\min}$, respectively. The coordinate values of the keypoints prior to normalization are x_i and y_i, respectively. The keypoints' level of confidence is C_i prior to normalization. The keypoints' horizontal and vertical coordinate values upon normalization are represented by the letters X_n and

Y_n, respectively. Following normalization, the error detection frame can be greatly reduced thanks to the keypoints' post-normalization confidence level, or C_n.

The normalized coordinate values are supplied into a Support Vector Machine (SVM) for training. The SVM classifier is highly fault-tolerant and recognizes skeletal differences even for the same behavior, preventing erroneous capture of CBs [16,17]. Moreover, SVM uses a Kernel Function, which is appropriate for high-dimensional issues and quick to train, to transfer linearly indistinguishable problems to a high-dimensional space. The flowchart of the capture algorithm design is presented in Figure 2 after training, which results in a model that captures the various classroom learning behaviors. The parameter "C" functions as a regularization, which represents the algorithm's capacity to fit nonlinearities, and is investigated here with a value of 1.0. The SVM was trained using the SVC function of the Scikit-learn package in Python. Kernel, the Gaussian Radial Basis Function, which is the default value for the "kernel" parameter, performs well for a variety of samples of any size.

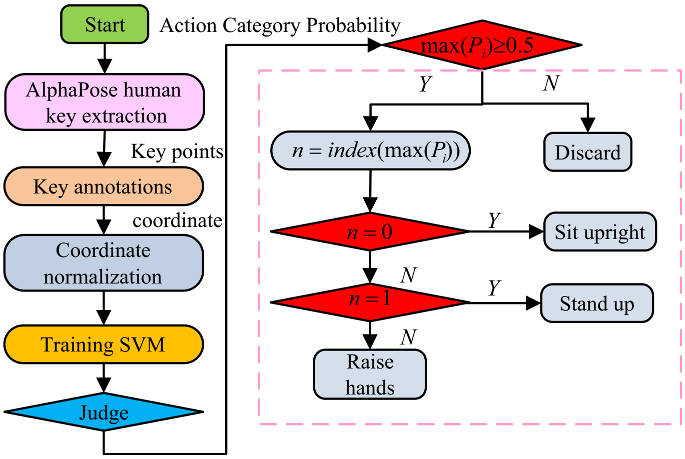

Figure 2: Student BC algorithm based on keypoints.

The single human detection frame is typically shaped like a vertical bar, but on occasion, there may be a mismatch between the critical points of two distinct targets, leading to an expanded skeleton and a horizontal detection frame. The model establishes a border threshold for filtering with a border length and breadth of 400 and 200, respectively, to prevent such misclassification. Equation (2) displays the filtering principles, where x and y stand for width and length, respectively.

$$\begin{cases} \max(x) - \min(x) < 200 \\ \max(y) - \min(y) < 400. \end{cases} \tag{2}$$

The extracted frames also contain a confidence score (proposal-score), and false detections outside the threshold need to be further set to continue to exclude false detections. The confidence threshold is set to 1, with a value greater than 1 indicating that the keypoint detection is complete and correct. Less than 1 means that the keypoint detection is incomplete. And the keypoint confidence is used as the training feature dimension, which can capture the CB under object occlusion condition.

The keypoint detection approach is more reliable and does not include background disturbance. The residual network (ResNet18) was chosen as the online student BC algorithm based on RGB picture categorization since the background knowledge can still be useful to BC. Early neural networks had a "simple-complex-simple-complex" structure, and when more layers were added, the neural network's capacity for learning increased. However, the neural networks did not show better performance as the number of layers increased, and degradation occurred, resulting in slower convergence of the neural networks. To address the problem of network degradation, the researcher proposes to use ResNet to solve it. The residual structure adds a constant function to the fitting function to keep the input and output consistent, as shown in equation (3), where $H(x)$ is the constant mapping function and $F(x)$ is the residual function.

$$\begin{cases} H(x) = x \\ H(x) = F(x) + x \\ F(x) = H(x) - x. \end{cases} \tag{3}$$

The gradient disappearance issue brought on by the Sigmoid activation function is likewise addressed by the ResNet network using the ReLU activation function [18,19]. ReLU is an unsaturated activation function whose mathematical equation is given in equation (4), where x denotes the input, and which solves the gradient disappearance issue while accelerating convergence.

$$y = \max(x, 0) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{else.} \end{cases} \tag{4}$$

The study uses the residual network as an image feature extraction network as it also simplifies the learning objectives and difficulty to some extent. The BC algorithm with RGB-based image classification first obtains a rough detection frame and labels it by the YOLO v3 target detection method. The same screening is performed to exclude incomplete and fuzzy detection data, and the data are pre-processed and input into ResNet18 for training, capturing the algorithm model as shown in Figure 3.

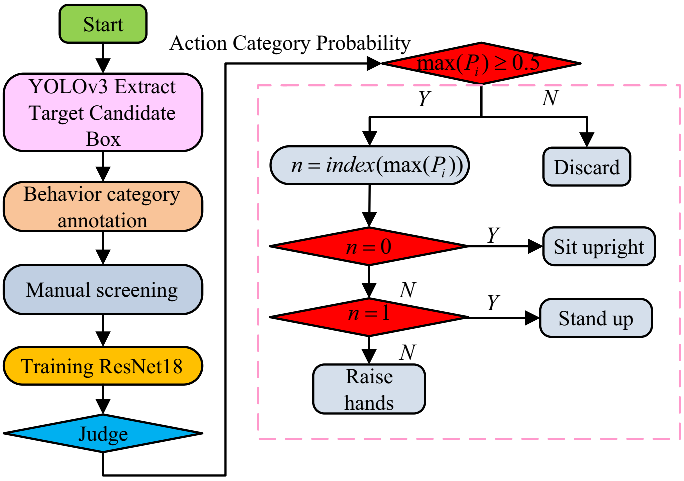

Figure 3: Classroom student BC algorithm based on RGB image classification.

3.2 Design of a dual flow BC and analysis system for fusion data

A single data type is less reliable for CB collection since it cannot access hidden information and can be interfered with background noise. In order to construct a dual flow network building block BC and analysis system, keypoint and RGB image data are combined in the study. Moreover, two branches are created to conduct BC simultaneously, drawing inspiration from the dual flow approach, and fusion is carried out using the channel connection Concat operation. The algorithm design flow chart is displayed in Figure 4.

The ResNet18 base network is based on migration learning for model training. ResNet18 uses the YOLO v3 target detection method to obtain a single person detection frame, and there are duplicate candidates. The study uses a non-maximum suppression method with improved head and shoulder regions for duplicate frame filtering, and after obtaining the keypoint coordinates of the head and shoulder regions, the maximum horizontal and vertical coordinate values are used as the length and width of the head and shoulder detection frames. A factor of 0.1 is also extended to ensure detection integrity. Intersection over union (IoU) is the ratio of

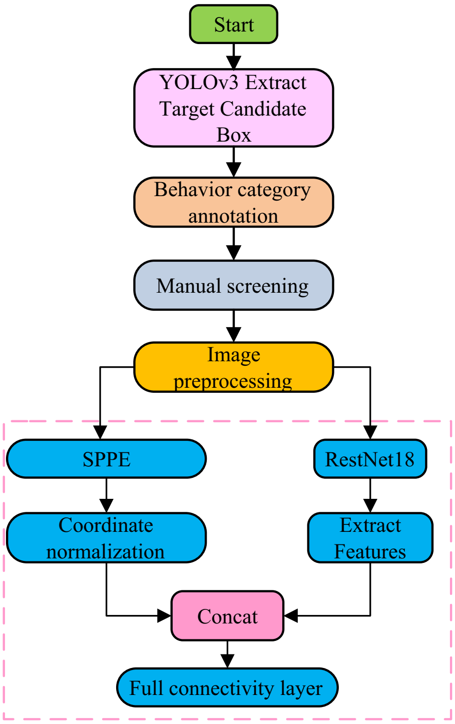

Figure 4: Classroom student BC algorithm based on keypoints and RGB images.

the candidate frame to the original marker frame, ideally 1 in the case of complete intersection, and the length, width, and center point coordinates of the required head and shoulder frame are calculated. The IoU value is filtered when it is greater than the study set threshold of 0.1.

Concert operation for feature fusion can expand the number of channels [20,21]. Equation (5) illustrates the operation process. X_i and Y_i stand for the two spliced channels in equation (5). The convolution operation is denoted by *, while the shared convolution kernel is denoted by K_i. Figure 5 depicts the entire procedure.

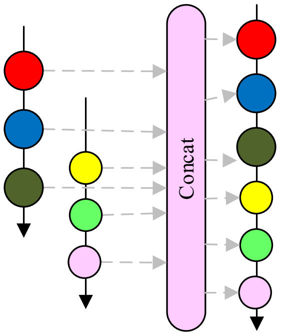

Figure 5: Concat operation diagram.

$$Z_{concat} = \sum_{i=l}^{c} X_i \times K_i + \sum_{i=l}^{c} Y_i \times K_{i+c}. \tag{5}$$

Equation (6) uses the squared loss function as the objective function for training the network model. Equation (6) illustrates the network optimization process using the stochastic gradient descent technique.

$$\begin{cases} L(Y, f(x)) = (Y - f(x))^2 \\ J(\theta) = \min\limits_{\theta} \dfrac{1}{2} \sum\limits_{i=1}^{m} (h_\theta(x^{(i)} - y^{(i)}))^2 \\ \theta_i = \theta_i - \alpha \dfrac{\partial}{\partial \theta_i} J(\theta), \end{cases} \tag{6}$$

where θ_i represents the value of the loss function and α represents the learning rate.

Before the model is trained, the photos are additionally pre-processed, first by removing images from the outside of the classroom and then by noise reduction of the image frames. Bilateral filtering is used for noise filtering to maximize the retention of boundaries while reducing noise at the time of taking the integrity of the image data into consideration. The weight in the spatial domain is G_s, and the weight in the pixel range is G_r. The calculation equation is shown in equation (7), where p represents a point and B represents the center of the window.

$$\begin{cases} G_s = \exp\left(-\dfrac{\|p - q\|^2}{2\sigma_s^2}\right) \\ G_r = \exp\left(-\dfrac{\|I_P - I_q\|^2}{2\sigma_r^2}\right). \end{cases} \tag{7}$$

The result of the bilateral filtering is shown in equation (8), where W_q is the sum of the weights obtained by summing the weights of each pixel of the corresponding filter window.

$$\mathrm{BF} = \frac{1}{W_q} \sum_{q \in S} G_s(p) \times I_p = \frac{1}{W_q} \sum_{q \in S} \exp\left(-\frac{\|p - q\|^2}{2\sigma_s^2}\right) \exp\left(-\frac{\|I_P - I_q\|^2}{2\sigma_r^2}\right) \times I_p. \tag{8}$$

Data enhancement is used to alter the geometric position of pixels in the image after denoising to create a new image. The categorization of the capture classification model can be improved by expanding the dataset. The two primary operations of enhancement are coordinate change and grey-scale interpolation, and equation (9), which is the spatial transformation coordinate employed in the study, contains the equation for the affine transformation's spatial coordinates. The original image's pixel coordinates are (v, w) in equation (9) and the modified image's pixel coordinates are (x, y).

$$\begin{cases} (x, y) = T\{(v, w)\} \\ [x \quad y \quad 1] = [v \quad w \quad 1]T = [v \quad w \quad 1] \begin{bmatrix} t_{11} & t_{12} & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & 1 \end{bmatrix}. \end{cases} \tag{9}$$

Finally, in an attempt to apply the model to a real system, Class Activation Maps (CAM) are used to visualize the error judgement data and to assist in understanding the learning of the dual flow network. CAM uses a migration learning approach where the convolutional layer is frozen fixed and the classification probability of the model only changes with the weight matrix W of the fully connected layer. During model training, the weight matrix W is saved for each update and summed up by dotting the weight matrix with the set of feature maps at different iterations in different batches. However, W does not visually reflect the learning of the features by the model, and a weighting operation needs to be performed to superimpose the new feature maps. The operation process is shown in Figure 6.

For an image, $f_k(x, y)$ denotes the activation value of the image at the coordinates (x, y) for cell k after the last convolution layer. The unit k is pooled on average as $\sum_{x,y} f_k(x, y)$ and then subjected to a dot product operation with W_k^c, the operation process is shown in equation (10). Where c denotes the different categories, M_c is the mapping of weights by performing a dot product operation of the weight matrix W_k^c with the feature map set $f_k(x, y)$.

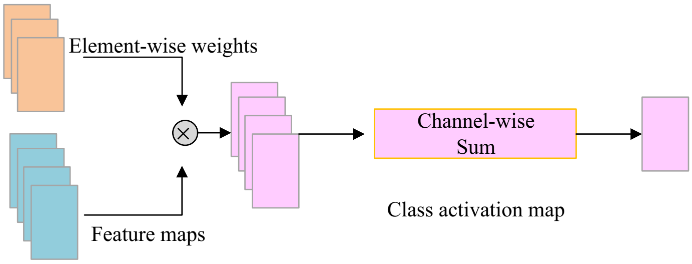

Figure 6: Class activation diagram.

$$\begin{cases} S_c = \sum_x W_k^c \sum_{x,y} f_k(x,y) \\ S_c = \sum_{x,y} \sum_k W_k^c f_k(x,y) \\ M_c(x,y) = \sum_k W_k^c f_k(x,y) \end{cases} . \tag{10}$$

4 BC and analysis system performance testing

4.1 Experimental scheme design and network parameters analysis

A test experiment was created to confirm the effectiveness of the system built for this study. The experimental hardware environment is a PC with a Core i5-5350U CPU @ 1.8 GHz, 16.0GB of memory, and NVIDIA GeForce GTX1070 GPU. The experimental algorithm is run on the Win7 operating system and Python platform based on x64. The dataset comes from video image recordings of the learning process retained from a large OT platform. All images are first filled with rectangular boxes to make them equal in length and width, and then their pixels are scaled to 256 × 256, and cropped to a uniform 224 × 224 size. After the data preprocessing operation, the dataset contains 412,640 images of all behaviors and the validation set contains 7,521 images. In order to ensure that the model learns image features better and avoid over-fitting phenomenon, data enhancement technique is used to expand the image data set by horizontal or vertical flipping, rotating, scaling, cropping, and shifting operations, and the final images that enter the training set after augmentation include 431,138 images, and the validation set contains 10,425 images. Before entering the network training, the image pixels and ImageNet mean values are subjected to the subtraction operation.

The parameters of the algorithmic model itself have a significant impact on the model's capacity for learning, and there are numerous elements influencing the training effect of the model. The accuracy rate is utilized as the evaluation metric, and the number of iterations is set at 100. On the performance of the model BC analysis, the effects of the learning rate α and the quantity of data batch size provided to the program for training in a single iteration are discussed. Keeping all other parameters constant, the impact of the learning rate on the model's accuracy is first addressed, with α set to 0.0001, 0.0005, 0.001, 0.005, and 0.05, respectively. The training results are displayed in Figure 7.

Figure 7 illustrates how the model accuracy curve roughly tends to rise as the number of iterations increases when the learning rate is between 0.0001 and 0.001. When the iterations are finished, the accuracy of the model with a learning rate of 0.001 increases by 8.1% from the learning rate of 0.0001 to a maximum of 92.3% for the same number of iterations. The accuracy curve, on the other hand, declined towards the lower accuracy region and had a maximum drop of 23.6% as the learning rate value increased. Particularly noteworthy is the fact that there is a significant decline in accuracy at the beginning of an iteration when the learning rate reaches a value of 0.01. Finally, a non-convergence outcome was obtained with the optimal learning rate selected at a value of 0.001.

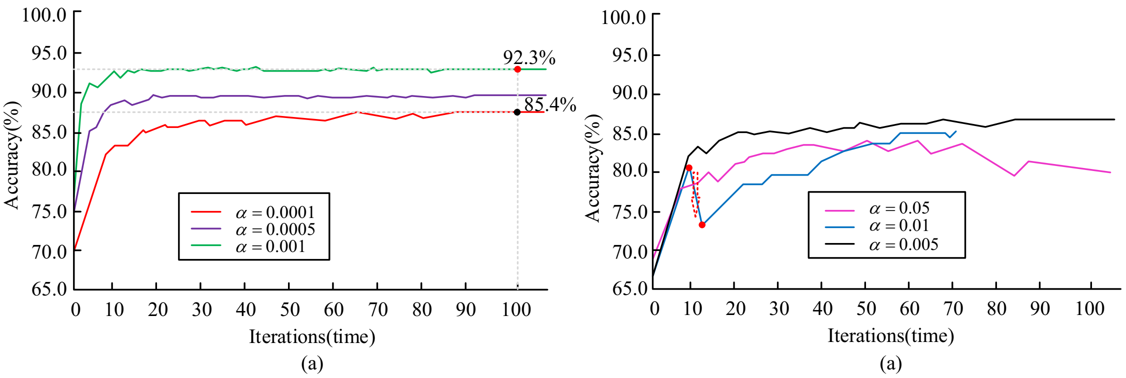

Figure 7: Comparison of accuracy of models under different learning rates. (a) 0.0001, 0.0005, and 0.01, and (b) 0.05, 0.01, and 0.005.

Batch size indicates the number of data passed to the program for training in a single pass, which has little effect on the training results when the number of samples is small. However, this experiment was performed with data augmentation operations, which may lead to memory explosion. Therefore, the learning rate was set to 0.001 and 100 iterations, and the accuracy rate was also used as the evaluation index. And Batch size was set to 4, 8, 16, 24, and 32, respectively, and the training results are shown in Figure 8. The accuracy rate increased with the number of iterations in the range of 4–8, up to 90.8%, but decreased when the value of Batch size increased. The accuracy curve is closer when the values are 24 and 32, so the best Batch size setting is 8.

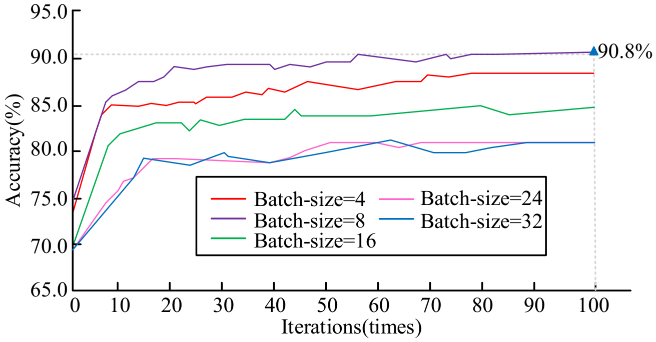

Figure 8: Comparison of accuracy of models under different batch size values.

4.2 System performance testing and analysis of results

Setting the model learning rate to 0.001 and the Batch size to 8, the model was trained on the chosen dataset. The research looked at how well the three behaviors – sitting, raising the hands, and standing – trained on various models. The two comparison models were built using, respectively, the RGB image classification and AlphaPose keypoint detection approach. Also, the experiments are evaluated for the training accuracy, precision, recall, and $F1$ values of the various models for the various behaviors.

The training accuracy of the three models is shown in Figure 9 after repeating the experiment ten times on the same dataset and calculating the standard deviation and standard error of the results. The accuracy rate is the ratio of the number of correctly predicted samples to the total number of samples among all the predictions of the model. The higher the accuracy rate, the better the model's predictive ability and the more accurate the BC. As seen in Figure 9, the data fusion algorithm has the highest accuracy curve and the highest average accuracy with a value of 97.3%. The accuracy of the RGB image classification-based algorithm has a large difference in accuracy with the other two algorithms, while the AlphaPose keypoint detection-based

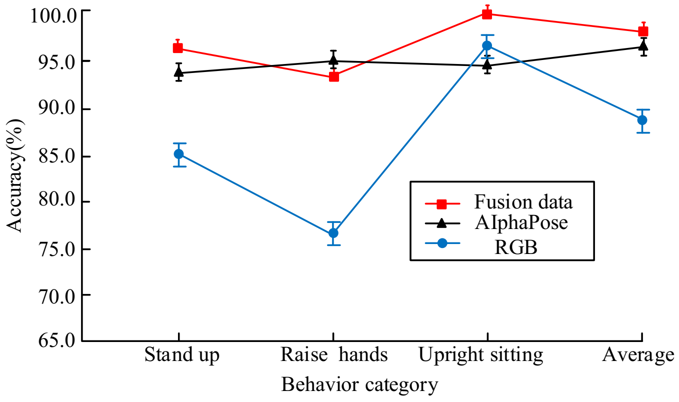

Figure 9: Accuracy of different algorithms for capturing different behaviors.

algorithm has a small difference in accuracy with the fused data algorithm. Thus, the dual flow network based on AlphaPose keypoint detection and fused RGB images on the surface is feasible for BC. Of particular note that the AlphaPose keypoint detection-based algorithm is 5.7 percentage points more accurate than the fused data algorithm for capturing hand-raising behavior, and the RGB image-based algorithm drops to 74.8% accuracy for capturing hand-raising. This result indicated that the RGB image-based algorithm was less capable of discriminating hand raising behavior, and therefore fused data still had a negative impact on the BC capability of the model.

The model was further examined for memory and precision, and the training outcomes for recall and precision are displayed in Figure 10. Precision rate is the proportion of the number of samples that are truly positive cases to the number of samples that are predicted to be positive cases out of the total number of samples that are predicted to be positive cases, and the precision rate measures how accurately the model predicts positive cases. Recall is the ratio of the number of samples predicted to be positive to the number of samples that are truly positive among all positive samples, and recall measures the model's ability to find positive cases. Often this pair of metrics is contradictory and can be combined to judge the model's accuracy. Figure 10 shows that the accuracy of the two-stream network using fused data is at the greatest level, demonstrating the algorithm's high accuracy and effective BC process performance. The RGB picture algorithm has the lowest accuracy rate for the CB of the ascending, uplifted hand. The noise interference in the image frames still had an impact on the algorithm's training results even after the image data had undergone noise reduction.

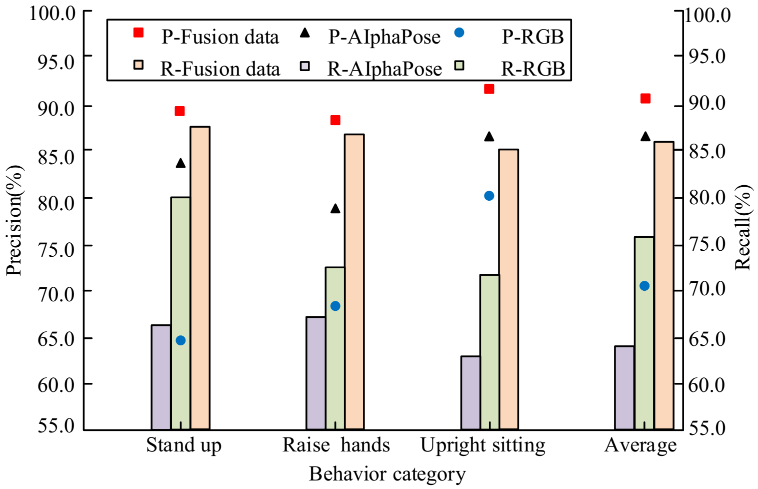

Figure 10: Accuracy and recall curve of different models.

The three algorithms were employed to record and analyze the various behaviors, and Table 1 contains the precise accuracy and recall rates of the algorithm training outcomes. Table 1 shows that adding the keypoint detection technique boosted accuracy by 25 percentage points for the rising behavior and by 19 percentage points for the hand raising behavior. The capture accuracy for the stand-up behavior was improved by 14 percentage points, and the capture accuracy for the three behaviors was improved by up to 10.6 percentage points compared to the traditional keypoint recognition technology, which is a significant improvement in the capture accuracy. In addition, precision and recall are two contradictory evaluation metrics, but it can be seen that the data fusion methods have basically the highest recall, with 88.3, 86.2, and 85.1% recall for capturing standing up, hand raising, and sitting down behaviors, respectively, which are in the range of 85% or more. While the recall of RGB image recognition technique is roughly in the range of 70–80%, the AlphaPose recognition technique is in the range of 60–68%. This shows that the data fusion algorithm does not maximize these opposing metrics by maximizing one while sacrificing the other.

Table 1: Results of different algorithms for three behaviors

Behavior classification	RGB		Data fusion		AlphaPose	
Evaluating indicator	**Accuracy (%)**	**Recall (%)**	**Accuracy (%)**	**Recall (%)**	**Accuracy (%)**	**Recall (%)**
Stand up	64.9	79.8	89.9	88.3	84.2	66.8
Rise hands	67.2	72.4	87.9	86.2	77.3	68.5
Upright sitting	78.4	71.6	92.4	85.1	85.6	62.4
Average	70.2	74.6	90.1	86.5	82.4	65.9

The *F*1 metric is a reconciled average of precision and recall, and is able to synthesize the model's performance on both positive and negative class samples. Precision and recall may be biased when the sample distribution is unbalanced, and *F*1 corrects the sample imbalance problem and makes it easy to interpret and compare the performance between different models, and has a maximum value of 1 and a minimum value of 0. The results of training the *F*1 value for the model to capture the telescoping behavior are shown in Figure 11. As seen in Figure 11, the *F*1 value for data fusion is the highest, with a maximum value of 93.1%, the *F*1 value curve is at the top of the curve, and the *F*1 values for the other two recognition techniques are only in the 85–90% range. Overall, it appears that the model performs optimally. When measured in terms of accuracy, precision, and recall, the performance of the RGB image-based algorithm for capturing behavior was not as good as the performance of the other two algorithms. This shows that this research is scientifically sound for CB capture and analysis based on keypoint construction and that the fused data algorithm has a degree of performance improvement over using one algorithm alone.

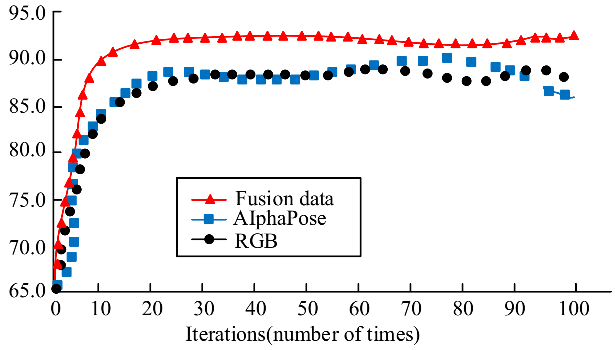

Figure 11: *F*1 values for different algorithm models.

Different categories of behaviors are set up with weights for constraints, and the weights are set according to the number of behavioral samples, and the sitting behavior has the largest number of samples and the smallest weight value. The recognition results of the two-stream network based on the weight settings are shown in Table 2. As can be seen from Table 2, changing the weight settings can further improve the recognition accuracy and the recall rate is kept at a comparable level.

Table 2: Comparison results of recognition rates of different behaviors after setting weights

Behavior classification	**[1.0, 1.0, 1.0]**		**[0.24, 1.0, 0.41]**	
Evaluating indicator	**Accuracy**	**Recall**	**Accuracy**	**Recall**
Stand up	0.97	0.94	0.99	0.95
Rise hands	0.75	0.91	0.84	0.92
Upright sitting	0.99	0.91	0.99	0.92
Average	0.90	0.92	0.94	0.93

5 Discussion

In the process of teaching, students' facial emotional expressions or classroom behavioral gestures are important teaching feedback. However, teachers' main focus is on teaching content, and students' CBs are complex and variable, which makes it difficult for teachers to fully observe students' CBs at the first time, which is not conducive to the quality of teaching and learning. Existing experts and scholars have carried out research on the supervision of CB, and the results show that AI-related technologies can help automate the analysis of CB, so that teachers can more efficiently and intuitively grasp the students' learning input, which helps optimize the teaching process and timely implementation of teaching interventions. Through OT platforms, learning management systems, and other auxiliary tools to collect students' classroom interaction behaviors, such as asking questions, answering questions, speaking, etc., after data cleaning, data conversion operations can be carried out to analyze learning behavior. The analysis of online CB can have multiple goals. On the one hand, by analyzing students' behavioral patterns and trends, we can understand students' engagement, learning interests, and learning styles, and provide a basis for personalized teaching and learning support. On the other hand, by analyzing the relationship between student behavior and academic performance, effective learning strategies and teaching methods are explored to provide teachers and educational decision makers with suggestions for improvement and optimization.

Based on this, the study investigates the algorithms for capturing and analyzing the behavior of students in online classrooms. The model designed in this research is first based on the HKD method to obtain the human behavior skeleton information, and then draws on the idea of dual stream method to construct a dual stream BC and analysis model by integrating the human keypoint extraction information with RGB images to achieve the purpose of identifying and analyzing CBs. The model can accurately identify common CBs such as raising hands, sitting up, standing up, etc. Adjusting the model parameter settings can further optimize the capture and analysis accuracy of the model. Compared with Pang's study, the proposed model not only utilizes the recognition of the human skeleton model, but also integrates the RGB images together for analysis, which avoids the errors caused by the overlapping and occlusion of the skeleton of the behavioral actions. Compared with Pang's clustering-random forest classification algorithm, the recognition accuracy of the research-constructed model is improved by 21.28%; compared with Pang's network topology model, the research-designed dual-stream network fusion is substantially reduced by 34.47% [22]. Compared with Wu's particle swarm-k proximity algorithm, the research constructed model minimizes the impact of simple image recognition susceptible to background noise interference, the highest behavior recognition accuracy reaches 97.3%, and the algorithm has superior comprehensive performance; compared with the key frame-based feature extraction, the human body key point detection method designed by the research greatly improves the algorithm's processing speed, and reduces the computational complexity [23].

The model designed by the study has been innovated in algorithmic technology, the algorithm based on the keypoint improvement improves the traditional recognition technology due to the background interference caused by the problem of recognition effect degradation, dual-channel data fusion improves the recognition accuracy; at the same time, the manual feature extraction is transformed into the human body posture estimation, and the recognition effect is further improved. The experiment successfully verified the practicality of the model, and the results of the research with other experts and scholars consistently showed that the research on the recognition of CB occupies an important position in the reform of education and teaching. The research results have been put into practical use in an online college English classroom in a domestic university, and the teachers reacted to the classroom students' behavior on a timely basis and accurately, grasped and understood the students' CB and teaching feedback in real time, and this novel teaching behavior analysis model also mobilized the students' motivation to learn, and the teaching atmosphere and quality of the whole classroom have been significantly improved compared with the previous ones.

6 Conclusion

Aiming at the technical difficulties of capturing and analyzing student behavior in online classroom, the study designed a dual-stream network integrating keypoint detection and RGB images for student BC. The experimental results show that the optimal learning rate of the network parameters is 0.001, at which time the model accuracy is 92.3%, an increase of 8.1% compared with the learning rate of 0.0001; the optimal Batch size is 8, at which time the model accuracy is 90.8%. The dual-stream network shows optimal performance in accuracy, precision, recall, and $F1$ value evaluation metrics, and the fused data algorithm achieves 97.3% accuracy in capturing end-sitting behaviors, but the accuracy of hand-raising behavior capture decreases by 5.7 percentage points, and the noise of the RGB image has an impact on the model performance. The accuracy of the dual-stream network incorporating the keypoint detection technique is at the highest level, with 25 percentage points improvement in the capture accuracy of the rising behavior and 19 percentage points improvement in the capture accuracy of the hand-raising behavior. The data fusion approach maximizes the contradictory metrics of precision and recall, with a maximum $F1$ value of 0.931 and good overall performance. The dual-stream network designed in the study effectively combines the advantages of the two techniques and improves the BC precision and the robustness of the algorithm. The model facilitates teachers to capture students' CB data in real time, assists teachers to regulate students' CB, improves students' learning status, and facilitates teachers' teaching management work. This study effectively overcomes the shortcomings of manual supervision in online teaching, and is of great significance in promoting the theoretical and practical reform of online teaching mode. However, the study only considered the fusion of keypoint detection and RGB data, and examined the capture and analysis of students' physical behaviors, while CBs containing richer semantic information, such as students' classroom speech and students' facial expressions, still need to be detected. The focus of future research could be to introduce computer vision technology to achieve the localization of classroom behavioral actions and to achieve more intelligent capture and analysis of online CBs.

Funding information: The research is supported by in 2021, Hebei Provincial Department of Science and Technology supported the project: "I'm coming for the Winter Olympics of Science and Technology" popular science book creation. Project No.: 20557604k.

Author contributions: The author confirms sole responsibility for the following: study conception and design, analysis and interpretation of results, and manuscript preparation.

Conflict of interest: The author declares no conflict of interest.

Ethical approval: Not applicable.

Data availability statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] Liu T, An Q, Huang Z, Xiong H, Cucchiera R, Deng Q. Efficient infrared imaging-blur kernel estimation with multi-scale feature learning for online learning video source. Infrared Phys Technol. 2022;434(120):103979–84.

[2] Yee BC, Nawi AM, Abdullah T. Potential disruptive innovation: online learning of public speaking courses in higher education. Foresight: J Future Studies, Strategic Think policy. 2022;24(3):445–55.

[3] Yu J, Oh H, Kim M, Jung S. Unusual insider behavior detection framework on enterprise resource planning systems using adversarial recurrent autoencoder. IEEE Trans Ind Inform. 2022;18(3):1541–51.

[4] Zheng Y, Zhang S. Research on fall detection based on improved human posture estimation algorithm. J Instrum. 2021;8(4):18–33.

[5] Lu M, Hu Y, Lu X. Pose-guided model for driving behavior recognition using keypoint action learning. Signal Process Image Commun. 2022;91(100):116513–20.

[6] Xiao W, Liu H, Ma Z, Chen W. Attention-based deep neural network for driver behavior recognition. Future Gener Comput Syst. 2022;48(132):152–61.

[7] Shi X, Huang J, Huang B. An underground abnormal behavior recognition method based on an optimized alphapose-ST-GCN. J Circuits Syst Comput. 2022;31(12):649–61.

[8] Zhang N, Yan Y, Zhu X, Wang J. A novel user behavior prediction model based on automatic annotated behavior recognition in smart home systems. China Commun. 2022;19(9):116–32.

[9] An J, Cheng Y, He X, Gui X, Wu S, Zhang X. Multiuser behavior recognition module based on DC-DMN. IEEE Sens J. 2022;22(3):2802–13.

[10] Wu L, Li X, Zhong K, Li Z, Wang C, Shi Y. HCCG: Efficient high compatibility correspondence grouping for 3D object recognition and 6D pose estimation in cluttered scenes. Measurement. 2022;197(57):111296–312.

[11] Zhu Z, Xiang W, Huo J, Yang M, Zhang G, Wei L. Non-cooperative target pose estimation based on improved iterative closest point algorithm. Syst Eng Electron Technol. 2022;33(1):1014–24.

[12] Bilal DK, Unel M, Tunc LT, Gonul B. Development of a vision based pose estimation system for robotic machining and improving its accuracy using LSTM neural networks and sparse regression – ScienceDirect. Robot Comput-Integr Manuf. 2022;74(1846):102262–89.

[13] Lin CJ, Shih CH, Wei TS, Liu PT, Shih CY. Local object tracking using infrared array for bed-exit behavior recognition. Sens Mater: An Int J Sens Technol. 2022;34(3):855–70.

[14] Jung M, Lee S, Sim ES, Jo MH, Lee YJ, Choi HB, Kwon J. Stagemix video generation using face and body keypoints detection. Multimed Tools Appl. 2022;81(27):8531–38542.

[15] Liu SC, Wang T, Zhang Y, Zhou R, Dai C, Zhang Y, Lei H, Wang H. Rethinking of learning-based 3D keypoints detection for large-scale point clouds registration. Int J Appl Earth Obs Geoinf. 2022;90(112):564–52.

[16] Wu Y, Li S. Damage degree evaluation of masonry using optimized SVM-based acoustic emission monitoring and rate process theory. Measurement. 2022;45(190):110729–44.

[17] Khan M, Reza MQ, Salhan AK, Sirdeshmukh SP. Classification of oils by ECOC based multi-class SVM using spectral analysis of acoustic signals. Appl Acoust. 2021;183(3):108273–84.

[18] Zan P, Zhong H, Zhao Y, Riu H, Dong R, Ding Q, Yue J. Research on improved intestinal image classification for LARS based on ResNet. Rev Sci Instrum. 2022;93(12):124101–16.

[19] Jung H, Rhee J. Application of YOLO and ResNet in heat staking process inspection. Sustainability. 2022;14(23):15892–903.

[20] Saha P, Neogy S. Concat_CNN: A model to detect COVID-19 from chest x-ray images with deep learning. SN Comput Sci. 2022;3(4):305–19.

[21] Chai E, Ta L, Ma Z, Zhi M. ERF-YOLO: A YOLO algorithm compatible with fewer parameters and higher accuracy. Image Vis Comput. 2021;116(45):104317–31.

[22] Pang C. Simulation of student classroom behavior recognition based on cluster analysis and random forest algorithm. J Intell Fuzzy Syst: Appl Eng Technol. 2021;40(2):24241–31.

[23] Wu S. Simulation of classroom student behavior recognition based on PSO-kNN algorithm and emotional image processing. J Intell Fuzzy Syst: Appl Eng Technol. 2021;40(4):7273–83.