Eng Int Syst (2025) 2: 201–211

Engineering
Intelligent Systems

Student Concentration Recognition Model in English Classroom Based on ResNet50 and VGG16

Dongdong Tang[1]* **and Qingqing Jiang**[2]

[1] *Qingdao Huanghai University, Qingdao 266427, China*
[2] *Haidu College, Qingdao Agricultural University, Yantai 265200, China*

Students' classroom behavior can indicate their listening status in real time, and teachers judge students' listening status based on their own experience and students' behavior. However, this method greatly tests teachers' teaching experience and energy, and it is difficult to achieve good results on this basis. Therefore, this study proposes an English classroom student focus recognition model based on deep residual networks and visual geometry group networks. The model selects the "You Only Look Once V3" algorithm as the target position detector, and then uses deep residual networks and visual geometry group networks to recognize and classify student classroom behavior, thereby determining each student's class state. The experimental results show that the accuracy of the cropped model is significantly higher than that before cropping. When the number of iterations reaches 500, the accuracy of the model before and after image cropping in the visual geometry group network model is 0.88 and 0.97, respectively. In regard to the deep residual network model, the accuracy of the model before and after image cropping is 0.86 and 0.98, respectively. For the dual network hybrid model, the model accuracy before and after image cropping was 0.90 and 0.99, respectively. The research results indicate that the proposed dual network hybrid algorithm model has excellent performance in recognizing student state.

Keywords: Behavior recognition, transfer learning, ResNet50, VGG16, concentration

1. INTRODUCTION

With the acceleration of economic growth, computer technology has also made great breakthroughs, and deep learning for efficient classrooms has become a development trend [1]. In teaching and learning activities, students are the core part of education, and their performance in the classroom directly determines the quality of education [2]. A comprehensive classroom-based education consists of all students' classroom behaviors, i.e., each student's behavior is the smallest unit in the whole classroom, and the students' listening concentration is closely analyzed through the monitoring of students' behavior such as note-taking, playing with cell phones, talking to each other, looking to the right and to the left, and so on. After observing the students' behaviors, teachers' teaching methods can be adjusted to improve the effectiveness of lectures. In the classroom, it is difficult to effectively and accurately recognize the student state due to the limited energy of the teacher and the difficulty of applying the traditional model of machine learning [3]. Therefore, in this study, an English classroom student concentration recognition model is proposed based on a deep residual network and a visual geometric cluster network, which chooses YOLOV3 as the detector of target position, which ensures the rapid detection of the target in complex environments, and then recognizes and classifies the students' classroom behaviors, aimed at providing the teachers with an opportunity to recognize the students' state in the classroom so as to improve the educating quality. The innovative contribution made by this research is as follows. Firstly, a dual network fusion model of VGG16 and ResNet50 is proposed, which combines the stability of VGG16 with the residual learning advantage of ResNet50 to solve the problems of deep network degradation and gradient disappearance, and improve classification accuracy and

*Email of Corresponding Author: tangdongdong0507@163.com

generalization ability. Secondly, the use of transfer learning methods to solve the problem of small-scale student behavior datasets has improved recognition performance by utilizing pre-trained models. Thirdly, integrating YOLOV3 for object detection effectively reduces background interference and improves model accuracy. Finally, a complete classroom behavior recognition system was developed, covering data collection, preprocessing, and model training, suitable for automated classroom management. This paper has four main sections. The first section provides a brief description of previous research on students' concentration in the classroom. The second section describes the methodology applied in this current study. The results of the model application and analysis are presented in the third section. The fourth section concludes the paper and suggests directions for future research.

2. LITERATURE REVIEW

The listening state of the students plays an important role in classroom conduct. Rafique et al. found that teachers play a key role in nurturing society and paving the way for socio-economic development. Therefore, the research team proposed a strategy to provide automatic analysis of the teaching methods of the teachers in the classroom environment, which is accomplished by 3D CNN and Conv2DLSTM. The outcomes indicate that the proposed solution provides strong support for teachers' teaching techniques [4]. Fang et al. studied and proposed a face recognition scheme supported by residual neural network and additional angular edge loss in order to address the shortcomings of facial recognition techniques by distinguishing facial features and improving the recognition accuracy. The outcomes show that the scheme is able to detect and recognize faces consistently, and its recognition accuracy remains high in complex situations where, for instance, there are facial defects and bright light exposure [5]. Pabba et al. addressed student disengagement in modern scenarios by proposing a model to monitor and analyze student engagement and behavior in real time. The model enables the recognition of student states in the classroom by analyzing their facial expressions and emotional states. The method can effectively monitor student engagement and behavior with high accuracy [6]. Tian et al. proposed a method supported by an improved anomaly detection model aiming to achieve more accurate facial expression recognition in a classroom environment. The method used an unsupervised learning approach of deep learning and constructed a face recognition model by means of a convolutional neural network. The model achieves 72.4% accuracy in expression recognition, with better classification accuracy [7].

Fan et al. proposed a hierarchical scale network to reinforce the accuracy of facial expression recognition. The method enhances the information. The outcomes indicate that the proposed solution exhibits high accuracy when applied to several types of data and the effectiveness was demonstrated by an ablation study [8]. Hu et al. found that the incidence of autism worldwide is increasing. In order to recognise the emotions associated with autism in children, a classroom emotion recognition model was proposed. The model can recognize children's expressions through a spatio-temporal graph convolutional network. The outcomes showed that the system can effectively determine the classroom emotions of children with autism and help educators [9]. Liu et al. proposed a blind image recovery solution using full variational regularization in order to improve the recognition of facial expressions in infrared video. The method is able to reveal the difference between low and high-resolution face expressions. The experiment outcomes suggested that the solution is able to recover high-resolution face images, which is beneficial for facial expression recognition [10].

In summary, many scholars have studied behavior recognition and achieved certain results, but no scholars have introduced action recognition into classroom teaching. A solution is put forward for recognizing students' concentration in English classroom based on deep residual network and visual geometric swarm network, which introduces action recognition into students' classroom, and judges students' classroom state by recognizing their actions, thereby allowing educators to improve the quality of classroom teaching.1.

3. CONSTRUCTION OF STUDENT CONCENTRATION RECOGNITION MODEL BASED ON RESNET50 AND VGG16

In this section, a database is constructed for students' classroom behavior, and the database is preprocessed. A transfer learning algorithm is proposed to address the problem of having a small dataset. Section II proposes a classroom behavior recognition model based on a dual network algorithm to recognize students' classroom state.

Most of the research in behavioral recognition is based on video and only very few studies are based on still images. When video technology was still in the developmental stage, information was always delivered through static images. In the classroom, students are the main body of learning and will engage in a variety of behaviors, which can usually be classified according to different states based on different classroom behaviors. Accordingly, students can be divided into attentive students and inattentive students, and their attention can be categorized as either college standard or inefficient classroom [11]. This study uses the S-T (Student-Teacher) method of student behavior analysis to represent the basic classroom behaviors through graphs, as shown in Figure 1.

In Figure 1, S denotes the student's behavior and T denotes the teacher's behavior. In the actual classroom, it is impossible for students to maintain the same posture, making it particularly difficult to select student actions, so the student posture above is selected in order to determine the subsequent action/ behavior. The images of students' behaviors are derived mainly from the network and video surveillance. Hence, before the dataset is constructed, it needs to be cleaned and filtered. Before the network experimental data is trained, the collected data needs to be enhanced [12]. Data augmentation is a technique whereby a series of transformations or pertur-

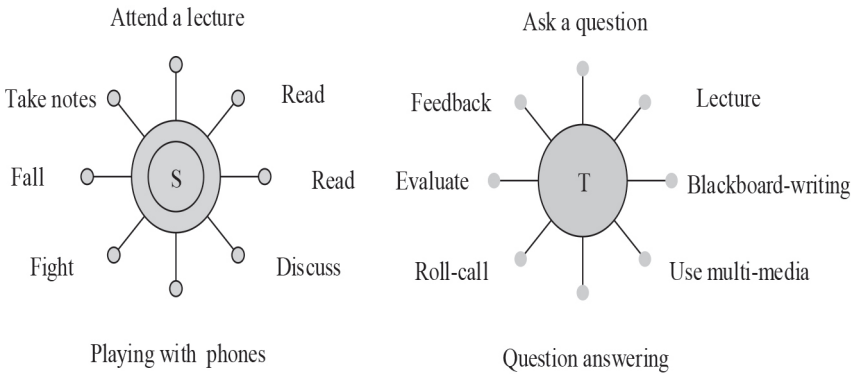


Figure 1 S-T behavior.

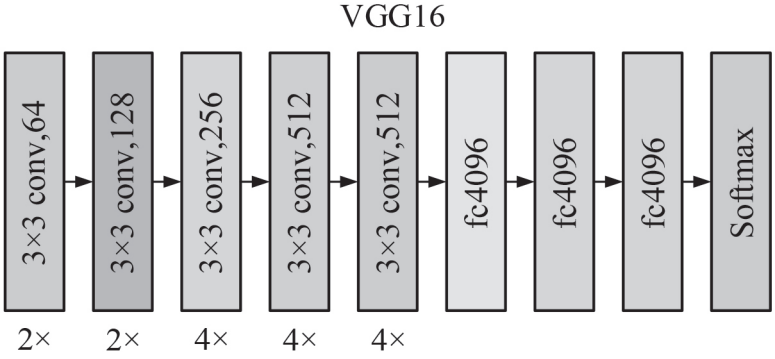


Figure 2 Diagram of VGG16 network.

bations are applied to the training data to create new training samples. This method helps to reinforce the generalization, mitigate overfitting and improve the performance. Some of the commonly used data enhancement methods are image panning, affine and perspective etc. Due to dataset limitations, migration learning methods are required for model training.

Transfer learning is a methodology that applies the knowledge learned from one task to another related task. Typically, machine learning obtains its required training with a large-scale dataset, and then the learned features, representations, or knowledge are applied to a new task, especially when the dataset for the new task is relatively small or has no annotation [13]. Visual Geometry Group Network (VGG Net) is a deep convolutional architecture proposed by Simonyan & Zisserman in 2014. It has a clean and deep structure, a smaller convolutional kernel and a deep network structure. Two common variants of VGG Net are VGG16 and VGG19, which contain 16 and 19 convolutional layers. The structure of the VGG16 chosen for this study is shown in Figure 2.

In Figure 2, VGG16 is a variant of VGG Net containing 16 convolutional and fully connected layers. The input size of VGG16 is shown in Equation (1) [14].

$$P = H_{in} \times W_{in} \times D_{in} \tag{1}$$

In Equation (1), P represents the input size, and H_{in} , W_{in} and D_{in} represent the height, width and depth of the input image, respectively. The output height of the convolution operation of the convolution layer is shown in Equation (2).

$$H_{out}=\frac{H_{in}-F+2P}{S}+1 \tag{2}$$

In Equation (2), H_{out} is the output image height, H_{in} is the input image height, F is the kernel size, P is the padding, and S is the size of the convolution step. The output width of the convolution operation of the convolutional layer is shown in Equation (3).

$$W_{out}=\frac{W_{in}-F+2P}{S}+1 \tag{3}$$

In Equation (3), W_{out} represents the width of the output image, W_{in} is the input image width, F is the kernel size, P represents the padding, and S is the size of the convolution step [15]. The pooling operation of the pooling layer is shown in Equation (4).

$$\begin{cases} H_{out} & = \frac{H_{in}-D}{H}+1 \\ W_{out} & = \frac{W_{in}-D}{H}+1 \end{cases} \tag{4}$$

In Equation (4), D represents size of the pooling window, and H represents the size of the pooling step size. The deep residual network (ResNet) is a deep neural network model used for image classification and target recognition. The traditional deep neural network learns the complexity of the input data layer by layer by stacking multiple layers. Hence, during training, it is prone to the gradient vanishing problem, making the model unable to converge [16]. The gradient problems are avoided by adding residual blocks to the neural network to allow the network to distinguish between features. ResNet consists of multiple residual blocks, each of which contains multiple residual units. The residual block output is shown in Equation (5) for an input value.

$$y = F(x, W_i) + xy = F(x, W_i) + x \tag{5}$$

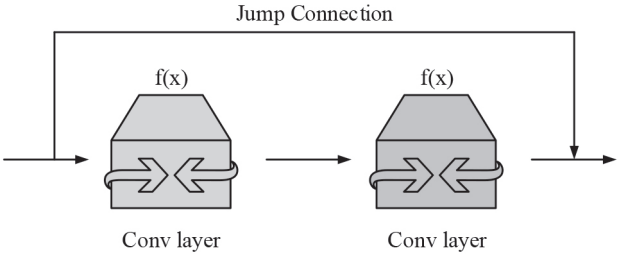


Figure 3 Structure of residual unit.

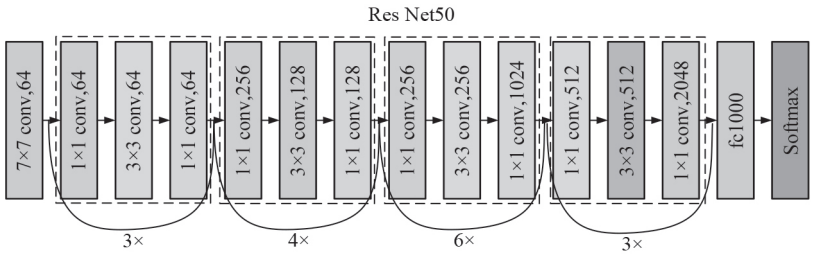


Figure 4 Structure of ResNet50 model.

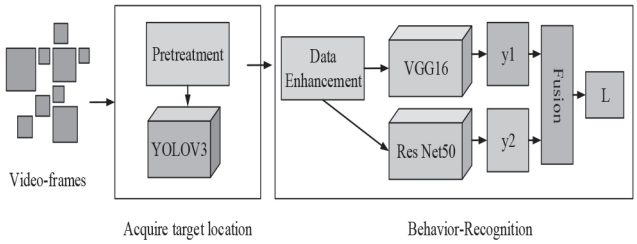


Figure 5 Classroom behavior identification model for students.

In Equation (5), x represents the input value, y is the output value, $F(x, W_i)$ denotes the mapping in the residual block, andW_i denotes the weight parameter of the residual block. The residual unit is the basic building block of ResNet. Jump connections are added between the two convolutional layers to map the input features directly onto to the output feature mapping [17]. The jump connections allow the forward propagation gradient to be passed directly to the later layers, preventing the gradient from vanishing. The residual unit is shown in Figure 3.

In ResNet, the batch normalization process and activation function are also used to enhance the expressive power. Equation (6) is used for batch normalization.

$$y=F(BN(x, W_i))+x \tag{6}$$

In Equation (6), BNdenotes the normalization process, whose activation function is ReLU, and the batch normalization process can accelerate the training process and bring up the performance and generalization ability. And normalizing the activations of the middle layer of the network on each training small batch helps to alleviate the problems of gradient vanishing and gradient explosion, while allowing the use of higher learning rate, which to some extent has the effect of regularization, and can mitigate the overfitting [18]. There are many models of ResNet, one of the most traditional being ResNet50, whose structure is shown in Figure 4.

As can be seen in Figure 4, ResNet50 has 50 convolutional layers, including a convolutional layer, a batch normalization layer, an activation function, and a fully connected layer. The input of the image is received at the input layer and the previous layers include convolutional and pooling layers for extracting the low-level features of the image. ResNet consists of multiple residual blocks. Each residual block contains multiple layers and functions. A global average pooling layer is used in the last layer of the network to transform the entire feature map into a vector. The fully connected layer is used to map the output of the pooling layer to the final classification result, with the output corresponding to a predefined number of categories.

3.1 Construction of a Dual Network Classroom Behavior Recognition Model

Feature extraction has evolved to become much finer, which requires the use of cascading ideas, which refers to a cascading processing or model combination idea in deep learning and computer vision [19]. This approach involves cascading multiple models or processing stages together to improve the performance, robustness or effectiveness of the system. This study uses a dual network algorithm model comprising YOLOV3 and VGG16-ResNet50, the structure of which is shown in Figure 5.

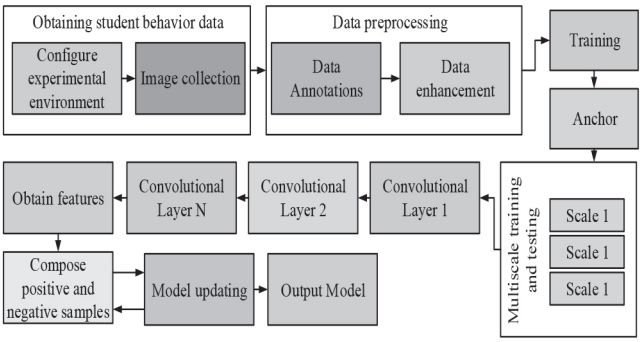


Figure 6 Training flow of YOLOV3 algorithm model.

As can be seen from Figure 5, the data are pre-processed and input into YOLOV3 for processing; then they are input into VGG16 and ResNet50 for feature classification output results. After the dataset has been constructed, the target location is detected through the YOLOV3 algorithm model, and then clipped to obtain a brand new set of data. In addition to the detection of the target, the background interference in the data also needs to be processed. In the training of students' classroom behavior data, the quality of the classifier determines the extent to which the features of the target object can be determined. The window sliding method is used to generate multiple candidate boxes on the image, and then the candidate boxes are subjected to feature extraction; the scores for each window are obtained after classification by the classifier [20]. The degree of convergence is judged by the change of the loss function. The convergence function is obtained with Equation (7).

$$Total\_loss=Giou\_loss+Conf\_loss+Prob\_loss \quad (7)$$

In Equation (7), $Giou\_loss$ represents the bounding box loss, $Conf\_loss$ is the symbol representing the confidence loss, and $Prob\_loss$ is the symbol representing the classification loss. Confidence loss is found with Equation (8).

$$Conf\_loss=bool*bce+(1-bool)*bce*ignore \quad (8)$$

In Equation (8), $bool$ denotes the confidence level, and bce denotes the binary cross-entropy loss of the center point coordinates. When there is insufficient original data, the dataset is enhanced and expanded by flipping, rotating and other operations performed on the dataset. The coordinate transformation of the dataset is obtained with Equation (9) [21].

$$(x, y) = T\{(v, w)\} \quad (9)$$

In Equation (9), (v, w) represents the original pixels coordinates and (x, y) represents the transformed graph. The affine transformation is a common transformation whose general form is shown in Equation (10).

$$\begin{bmatrix} x & y & 1 \end{bmatrix} = \begin{bmatrix} v & w & 1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & 1 \end{bmatrix} \quad (10)$$

In Equation (10), (v, w) represents original image pixels coordinates and (x, y) represents the transformed one. The YOLOV3 algorithm training is given in Figure 6.

As can be seen in Figure 6, firstly, the student's behavioral data is obtained and then preprocessed. The processed data is passed through Ancher to generate multi-scale data. Its features are obtained through convolutional layers to constitute positive and negative samples. The model is updated through the samples, and finally the model is output [22]. The features of the image are extracted using dual network fusion. ResNet50 can solve the problem of network degradation and the more difficult training. ResNet50 network structure contains a shortcut operation that enables the deep feature map to contain the information of the shallow feature map. The extraction of the target depth features can also avoid the problem of gradient dispersion, thus reducing the training error rate. VGG16 adopts the method of convolutional concatenation, which decreases parameters usage to make the feature extraction more accurate, and has a better adaptability to complex tasks. Moreover, the VGG16 model is relatively more stable and can be better combined with other networks. When the model is trained, the performance is indicated by the loss function. The cross-entropy loss function of ResNet50 is shown in Equation (11):

$$L=\Sigma_{i=1}^{M} y_i log(p_i(x)) \quad (11)$$

where L denotes the loss function, M denotes the total number of samples, p_i denotes the predicted probability of the observed samples, and y_i denotes the indicator variables. The cross-entropy loss function not only measures the similarity between the two, but also addresses the problem of learning rate degradation. The fused ResNet50 and VGG16 model is expressed with Equation (12):

$$y=\beta y_1+(1-\alpha)y_2 y=\beta y_1+(1-\alpha)y_2 \quad (12)$$

where β represents the output distribution of the VGG16 model, y_1 represents the weight ratio of the VGG16 model, α represents the output distribution of the ResNet50 model, and y_2 represents the weight ratio of the ResNet50 model. The hybrid model formed by integrating two types of networks can improve recognition accuracy and also enhance the stability of the model. The model process is as follows: firstly, a student behavior database is constructed and preprocessed, and YOLOV3 is used for object detection. The data processed by YOLOV3 is input into VGG16 and ResNet50 for feature extraction and classification, respectively. VGG16 performs deep feature extraction through convolution concatenation. ResNet50 contains residual blocks that can effectively alleviate network degradation through skip connections, improving

training stability and performance. The model is optimized using a loss function, and the classification results are measured by cross entropy. Combining the advantages of VGG16 and ResNet50, the model weights are fused to improve the classification accuracy and generalization ability of the model.

4. PERFORMANCE ANALYSIS OF STUDENT CONCENTRATION RECOGNITION MODEL BASED ON RESNET50 AND VGG16

The hardware configuration used in this experiment is Intel Core i5–8750H CPU, GPU is NVIDIA Geforce GTX2080Ti with 8 GB of video memory, and 16 GB of RAM. The dataset is a homemade dataset, which has eight types of students' images such as medium and medium note-taking, arguments, cell phone playing, discussions and lying on the table, totalling more than two thousand images. The performance of the algorithm on the dataset before and after cropping is compared. The results are shown in Figure 7.

Figure 7 (a) shows the recognition accuracy of the VGG16 algorithm model for images before and after cropping with different number of iterations, Figure 7 (b) shows the recognition accuracy of the ResNet50 algorithm model for images before and after cropping with different number of iterations, Figure 7 (c) shows the recognition accuracy of the VGG16+ResNet50 algorithm model for images before and after cropping with different numbers of iterations. Figure 7 (d) depicts the recognition accuracy curve for training using the cropped dataset. As shown in Figure 7, in the image data before and after cropping, the accuracy of the model after cropping is significantly stronger than that before cropping, and the accuracy of the model before and after image cropping is 0.88 and 0.97 in the VGG16 model under 500 iterations. The accuracy of the model before and after image cropping is 0.86 and 0.98 in the ResNet50 model, and the accuracy before and after image cropping is 0.86 and 0.98, respectively, in the VGG16+. ResNet50, the accuracy before and after image cropping is 0.90 and 0.99, respectively. Of the three methods, the VGG16+ResNet50 model has the best performance. The experimental results indicatethat the dual network algorithm has superior performance. The performances of the algorithms for the images before and after cropping are compared. The results are shown in Figure 8.

Figure 8 (a) shows the AUC values of the three models before image cropping. Figure 8 (b) shows the AUC values of the three models after image cropping. From Figure 8 (a), it can be seen that among the three methods, the area under the curve of the dual network hybrid algorithm is the largest; hence, of the three models, the dual network hybrid algorithm has the largest AUC value and the best performance. As seen in Figure 8 (b), after cropping the image, for each method, the area under the curve increases; the area of the dual network hybrid algorithm is the largest. The experimental resultsshow that the proposed dual network hybrid algorithm has the best performance. The recognition time of the algorithms is analyzed by modeling the three methods using different numbers of iterations. The results are shown in Figure 9.

Figure 9 (a) shows the recognition time of the three methods in the training set of different sizes, while Figure 9 (b) indicates the recognition time in the validation set of different sizes. As seen in Figure 9 (a), the recognition time required by each algorithmic model gradually decreases as the training set increases, and the model performance is gradually increasing. When the dataset is 1000, the recognition times of VGG16, ResNet50 and dual network algorithm models are 0.27 s, 0.35 s and 0.21 s, respectively. From Figure 9 (b), with the increase of validation set, the recognition times of each algorithm model increase, and when the validation set size is 500, the recognition times of VGG16, ResNet50 and dual network algorithm models are 0.41 s, 0.35 s and 0.21 s, respectively. Time is 0.41 s, 0.54 s and 0.28 s. The experiment outcomes said that the dual-network algorithmic model has better performance. The three algorithmic models are compared with the image data of students in eight different classroom states: head on the table, playing with cell phone, reading, raising hand, listening, arguing discussing, and taking notes, are selected for comparison, as Figure 10.

Figure 10 shows the recognition accuracy of the three algorithmic models for the eight students' classroom states, from the overall point of view, the area of the dual network model is larger than the other two algorithmic models. In terms of recognition actions, the recognition rate of the three methods for reading books is lower compared to other actions, indicating that all three models find it difficult to recognize the reading state; the average accuracy of the VGG16 algorithmic model, the ResNet50, and the dual network algorithmic model for each of the classroom states is around 80%, 75%, and 95%, respectively. The experimental results demonstratethat the proposed dual network algorithm model has better recognition accuracy for different class states. Two of the states with better recognition accuracies and two of the states with worse accuracies are selected for further analysis, namely, raising hands and listening, discussing and arguing. The results are shown in Figure 11.

Figure 11 (a) depicts the recognition accuracy of different method models and Figure 11 (b) shows the recognition time of different methods. As shown in Figure 11 (a), all the methods have low recognition accuracy for discussion and argument, and the VGG16 algorithm model, ResNet50 and dual network algorithm model have 68%, 72% and 86% recognition accuracy for discussion, and 70%, 74%, and 90% recognition accuracy for argument, respectively. Figure 11 (b) shows that three algorithms require longer recognition time for discussion and argument; the recognition time for discussion is 0.30 s, 0.28 s, 0.23 s, and for argument is 0.35 s, 0.32 s, 0.27 s, respectively. The experimental results indicate that the individual methods have poorer recognition effect for discussion and argument. The main reason is that the indicators of discussion and the argument are not easily recognizable, particularly because they have certain similarities. This means that it is easy to misjudge these behaviors when extracting local actions, which leads to failure to extract enough features. However, the overall recognition

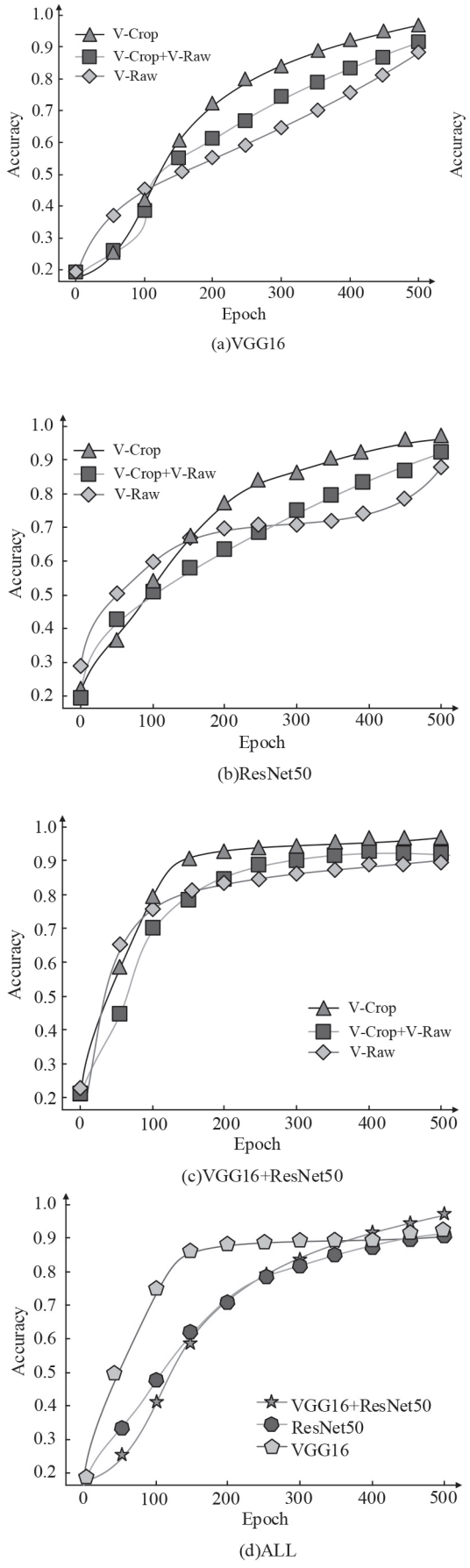
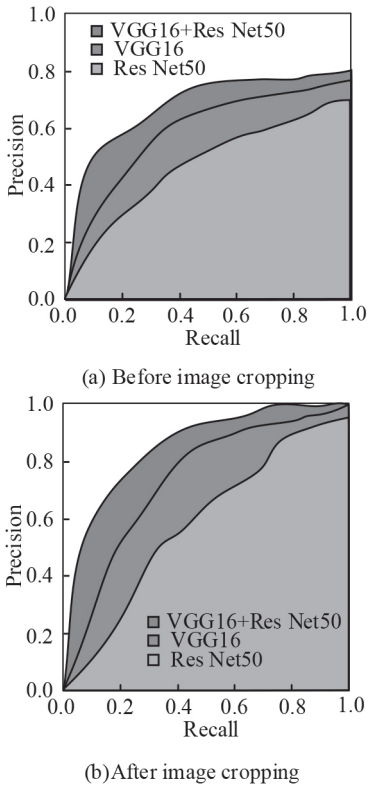


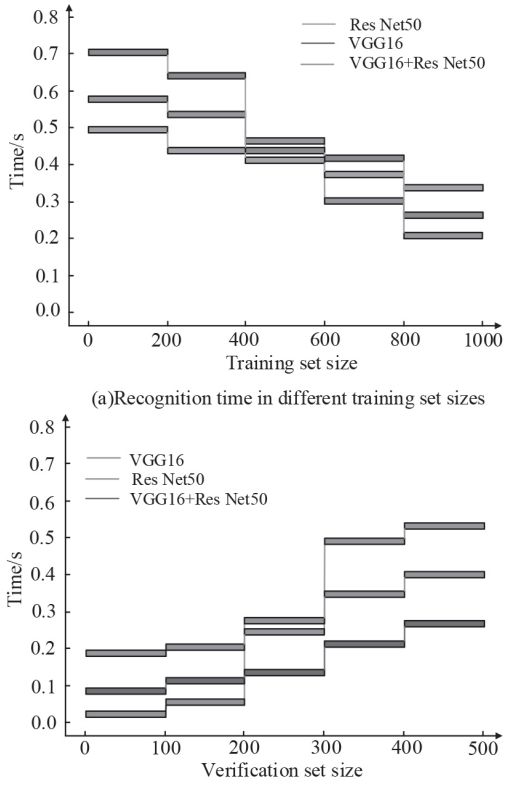
Figure 7 Comparison of recognition accuracy before and after image cropping.



(a) Before image cropping

(b)After image cropping

Figure 8 Comparison of model performance before and after image cropping.



(a)Recognition time in different training set sizes

(b)Recognition time in different verification set size

Figure 9 Comparison of model recognition time of the three algorithms.

accuracy is still high, which can meet the requirements of student behavior recognition in college classrooms. A certain classroom was selected and classroom analysis was conducted using three models, as shown in Table 1.

As seen in Table 1, the VGG16+ResNet50 model performs the best in terms of recognition accuracy, with an average of 95%, while the accuracy of VGG16 and ResNet50 is 85% and 80%, respectively. Especially when identifying

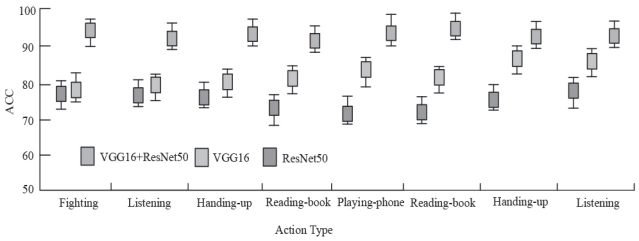
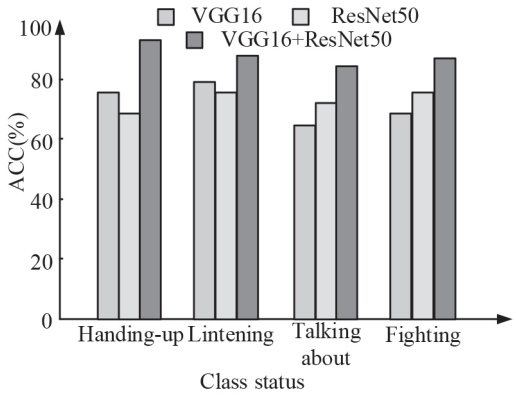
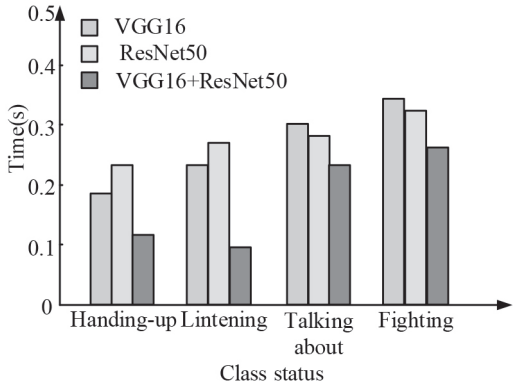


Figure 10 Recognition accuracy for different student states.



(a) The recognition accuracy of various student states



(b) Recognition time for various student states

Figure 11 Analysis of recognition performance for four class states.

Table 1 Analysis of model accuracy in recognizing students' classroom behavior.

Evaluation Metric	VGG16+ResNet50	VGG16	ResNet50
Recognition Accuracy (Average)	95%	85%	80%
Recognition Accuracy (Different States)	/	/	/
- Raising Hand	98%	92%	89%
- Taking Notes	95%	88%	85%
- Using Phone	90%	80%	75%
- Discussion	86%	68%	72%
- Arguing	90%	70%	74%
Recognition Time (Average)	0.28 s	0.41 s	0.54 s
AUC	0.99	0.95	0.92
Teacher Evaluation (Average Score)	89.9	82.6	78.5
Misrecognition Rate	5%	12%	15%
Training Time (Per Epoch)	30 min	25 min	27 min
Model Stability	High	Medium	Low

Table 2 Teacher evaluation form.

/	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teacher 5	Average
VGG16+ResNet50	92.1	90.7	87.6	85.4	93.7	89.9
VGG16	87.5	80.6	80.4	78.9	85.6	82.6
ResNet50	82.4	75.8	81.2	72.4	80.7	78.5

complex behaviors such as "arguing" and "discussion", the VGG16+ResNet50 model performs particularly well, with an accuracy rate of 86% for "discussion", significantly higher than VGG16's 68% and ResNet50's 72%. In terms of recognition speed, VGG16+ResNet50 takes 0.28 seconds, which is better than VGG16's 0.41 seconds and ResNet50's 0.54 seconds, demonstrating its advantage in real-time applications. In addition, the area under the curve of the dual network model reached 0.99, indicating its strong predictive ability and a false recognition rate of only 5%. The experimental results show that VGG16+ResNet50 is able to strike a good balance between accuracy, speed, and stability, making it the most suitable model for classroom behavior recognition. Fifty teachers were randomly selected and divided into five groups to evaluate the recognition performance of the model, as shown Table 2.

As shown in Table 2, the five groups of teachers gave the VGG16+ResNet50 model these ratings: 92.1, 90.7, 87.6, 85.4, and 93.7, with a mean score of 89.9. The ratings for the VGG16 model were 87.5, 80.6, 80.4, 78.9, and 85.6, with a mean score of 82.6. The ratings for the ResNet50 model were 82.4, 75.8, 81.2, 72.4 and 80.7, with an average score of 78.5. The experiment outcomes indicated that the proposed dual network hybrid algorithm model is preferred.

5. CONCLUSION

The students' listening state is a direct indication of the students' attentiveness. This study proposes a model for recognizing students' concentration in an English classroom based on deep residual network and visual geometric group network. YOLOV3 was chosen as the detector of target position, which can ensure the rapid detection of the target in acomplex environment. The experiment results showed that after 500 iterations, the accuracy of the model before and after image clipping in the VGG16 model is 0.88 and 0.97, respectively, and the accuracy of the model before and after image clipping in the ResNet50 model is 0.86 and 0.98, respectively. The accuracy of the model before and after image clipping in the VGG16+ResNet50 model is 0.90 and 0.99, respectively. 0.99. The recognition times for the VGG16, ResNet50 and dual network algorithm models are 0.27 s, 0.35 s and 0.21 s, respectively, when the training set is 1000. When the validation set size is 500, the recognition times for the VGG16, ResNet50 and dual network algorithm models are 0.41 s, 0.54 s and 0.28 s, respectively. The recognition times for the VGG16 algorithmic model, ResNet50 and dual network algorithmic model have recognition accuracies of 68%, 72% and 86% for discussions and 70%, 74% and 90% for fights, respectively. The experimental results indicate that the proposed VGG16+ResNet50 model is an effective means of recognizing students' state. However, this study has several shortcomings: that dataset is small and there is not enough variety in the adopted data. Hence, subsequent studies should choose a larger dataset to ensure a more adequaltey trained model.

FUNDING

This study was supported by phased research results of the horizontal research project "Student Concentration Recognition Platform in Business English Classroom from the perspective of Artificial Intelligence" of Qingdao Huanghai University (project host: Dongdong Tang).

REFERENCES

1. Zhang J, Wang X, Wan Y, Wang L, Wang J, Yu P, 2023, "SOR-TC: Self-attentive octave ResNet with temporal consistency for compressed video action recognition", Neurocomputing, vol. 533, no. 5, pp. 191–205.
2. Wei Y, Zeng A, Zhang X, Huang H, 2022, "RAG-Net: ResNet-50 attention gate network for accurate iris segmentation", IET Image Processing, 16, no. 11, pp. 3057–3066.
3. Xiao Y, 2024, "English learning behaviour pattern mining and personalized teaching strategies based on big data analysis", Engineering Intelligent Systems, vol. 32 no. 6, pp. 647–657.
4. Rafique M A, Khaskheli F, Hassan M T, Naseer S, Jeon M, 2022, "Employing automatic content recognition for teaching methodology analysis in classroom videos", Plos One, vol. 17, no. 2, pp. 546–556.
5. Fang J, Lin X, Tian J, Wu Y, 2022, "Face recognition technology in classroom environment based on ResNet neural network", Journal of Electronic Imaging, vol. 31, no. 5, pp. 5142–5153.
6. Pabba C, Kumar P, 2022, "An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition", Expert Systems, vol. 39, no. 1, pp. 154–162.
7. Tian J, Fang J, Wu Y, 2022, "Facial expression recognition in classroom environment based on improved Xception model", Journal of Electronic Imaging, vol. 31, no. 5, pp. 5141–5152.
8. Fan X, Jiang M, Shahid A R, Yan H, 2022, "Hierarchical scale convolutional neural network for facial expression recognition", Cognitive Neurodynamics, vol. 16, no. 4, pp. 847–858, 2022.
9. Hu B, 2022, "Analysis of art therapy for children with autism by using the implemented artificial intelligence system", International Journal of Humanoid Robotics, vol. 19, no. 3, pp. 53–73.
10. Liu X, Liu T, Zhou J, Liu H, 2023, "High-resolution facial expression image restoration via adaptive total variation regularization for classroom learning environment", Infrared Physics and Technology, vol. 128, no. 5, pp. 147–152.

11. Zhao W, Su Y, Hu M, Zhao H, 2022, "Hybrid ResNet based on joint basic and attention modules for long-tailed classification", International Journal of Approximate Reasoning, vol. 140, no. 10, pp. 83–97.
12. Fang J, Lin X, Tian J, Wu Y, 2022, "Face recognition technology in classroom environment based on ResNet neural network", Journal of Electronic Imaging, vol. 31, no. 5, pp. 51421.1–51421.16.
13. Zhou W, Han X, Xu Y, Chen R, Zhang Z, 2022, "Embryo evaluation based on ResNet with AdaptiveGA-optimized hyperparameters", Journal of Internet Technology, vol. 23, no. 3, pp. 527–538.
14. Yan Z, Liu H, Zhang S, Li J, Wang Y, 2022, "Superiority of two-dimensional correlation spectroscopy combined with ResNet in species identification of bolete", Infrared Physics and Technology, vol. 57, no. 23, pp. 579–581.
15. Yoo H N, Park M K, Park B G, 2023, "Effect of layer-specific synaptic retention characteristics on the accuracy of deep neural networks", Solid-State electronics, vol. 200, no. 1, pp. 108570.1–108570.6.
16. Thakur R, Panghal D, Jana P, Rajan, Prasad A, 2023, "Automated fabric inspection through convolutional neural network: an approach", Neural Computing &Amp, Applications, vol. 35, no. 5, pp. 3805–3823.
17. Li W, Li L, Yang H, 2023, "Progressive cross-domain knowledge distillation for efficient unsupervised domain adaptive object detection", Engineering Applications of Artificial Intelligence, vol. 119, no. 12, pp. 105–117.
18. Yan J, Wang Z, 2022, "YOLOV3+VGG16-based automatic operations monitoring and analysis in a manufacturing workshop under industry 4.0", Journal of Manufacturing Systems, vol. 63, no. 14, pp. 134–142.
19. Sarker S, Tushar S N B, Chen H, 2023, "High accuracy keyway angle identification using VGG16-based learning method", Journal of Manufacturing Processes, vol. 98, no. 5, pp. 223–233.
20. Shahabi M S, Shalbaf A, Nobakhsh B, Rostami R, Kazemi R, 2023, "Attention-Based convolutional recurrent deep neural networks for the prediction of response to repetitive transcranial magnetic stimulation for major depressive disorder", International Journal of Neural Systems, vol. 33, no. 2, pp. 235–247.
21. Pal S, Roy A, Shivakumara P, Pal U, 2023, "Adapting a swin transformer for license plate number and text detection in drone images", Artificial Intelligence and Applications, vol. 1, no. 3, pp. 145–154.
22. Stephen A, Punitha A, Chandrasekar A, 2023 "Designing self attention-based ResNet architecture for rice leaf disease classification", Neural Computing & Applications, vol. 35, no. 9, pp. 6737–6751.