Journal of Cases on Information Technology
Volume 25 • Issue 1

Novel Bilinear Fusion Network Based on Multimodal Data for Student Distracted Behavior Recognition: BFNMD

Jian Zhang, Zhengzhou Medical College, China*

ABSTRACT

As governments, education departments, and academic accreditation bodies have begun to encourage schools to develop evidence-based decision-making and innovation systems, learning analysis techniques have shown great advantages in decision-making aid and teaching evaluation. After integrating relevant algorithms and technologies in artificial intelligence and machine learning, learning analysis has achieved higher analysis accuracy. In order to realize the recognition of students' classroom behaviors such as standing up, sitting up, and raising hands and improve the recognition accuracy and recall rate, multi-modal data such as human key point information and RGB images are used for experiments. To further improve the feature extraction capability of the model, features are extracted from the improved ResNet-50 and EfficientNet-B0 models, and bilinear fusion is performed to further improve the recognition accuracy of the models.

KEYWORDS

Behavior Recognition, Bilinear Fusion, EfficientNet-B0, Multimodal Data

1. INTRODUCTION

Classroom is an important place for teachers to teach and students to acquire knowledge. With the continuous development of the society and the enhancement of the emphasis on student education, the intelligent analysis of classroom teaching quality becomes more and more important. Using information technology to detect, process and analyze students' behavior in class can not only remind students to standardize their behavior in class, but also reflect the active degree of class and help teachers improve teaching methods (Wu et al. 2020; Luo et al. 2015).

At the same time, in order to realize the rapid and extensive sharing of high-quality educational resources, video recording and broadcasting technology has been developed. Video recording and broadcasting system is a kind of educational system which uses multimedia technology to shoot and record classroom teaching activities in real time, and broadcast them live or on demand through the Internet (Meng et al. 2013). Traditional video recording and broadcasting system needs manual real-

DOI: 10.4018/JCIT.326131 *Corresponding Author

time shooting of teaching content, teachers in the classroom, blackboard writing, students stand up and sit down and other situations need to artificially control the camera to track the moving target. Therefore, extra manpower is needed to operate the camera, which leads to the instability of shooting quality and the increase of labor cost. In addition, the behavior of the filming staff controlling the camera and moving around in the classroom may interrupt the teacher's teaching ideas or distract the attention of the students, which affects the teaching quality to a certain extent.

With the development of artificial intelligence, deep learning and computer vision technology, it greatly promotes the application of intelligent video recording and broadcasting system, overcomes the shortcomings of previous manual monitoring, and has significant advantages in recognition performance, efficiency and other aspects. It only needs to install a camera in the classroom in advance, detect the behavior state of students in the classroom by using target detection and behavior recognition technology, and control the gimbal camera to track or shoot close-up pictures of students according to their state. The whole recording process does not require human participation, achieving a major breakthrough in video recording and broadcasting technology (Novakovsky et al. 2023; Wang et al. 2022).

However, there are few papers on classroom behavior recognition in academic circles, and the research methods mainly focus on machine learning and deep learning. (Cheng et al. 2022) obtained data from the number of faces, contour features and the range of subject actions, and used Bayesian causality network to deduce the subject behavior characteristics to identify students' behaviors. (Ahmad et al. 2008) extracted Zernike moment feature, optical flow feature and global motion direction feature of actions, and combined with naive Bayes classifier to recognize students' behaviors. The above method mainly uses traditional machine learning method, which requires tedious manual feature extraction steps and has low accuracy. (Jones et al. 2011) extracted the students' target area through background difference method and input it into VGG network China, and successfully identified three kinds of students' classroom behaviors: sleeping, playing mobile phone and normal.

The above references are all based on the data of a single mode for training and recognition, resulting in poor recognition effect of the model. Our main contributions are as follows. In this paper, through experimental comparison, Study the effect of multi-modal data on improving the accuracy and recall rate of classroom students' behavior recognition, and find that multi-modal data such as human key points, RGB The reasonable combined application of images and other information as well as the effective processing of data, such as re-labeling the data with large losses according to the ranking of training loss values, can effectively improve the effect of the recognition task and provide a new technical scheme for the recognition of students' classroom behavior.

This paper is organized as follows. Section 2 introduces the related works. Data acquisition and pre-processing are shown in section 3. Section 4 displays the proposed behavior recognition method. Experiments are conducted in section 5. There is a conclusion in section 6.

2. RELATED WORKS

Student engagement can help schools better understand the quality of student learning. The core factor to evaluate the education quality of a university is the degree of students' learning engagement. As an important part of learning engagement, students' classroom behavior has always attracted the attention of researchers. Traditional classroom behavior evaluation of students is achieved by manual observation records, which is inefficient. With the vigorous development of artificial intelligence today, attempts are made to improve this situation with the help of artificial intelligence technology (Groccia 2018; Jisi et al. 2021). To understand the learning behavior and state of students in the classroom learning process has become an important issue in the development of education, which will promote the intelligent, efficient and comprehensive development of educational analysis system. In order to promote the innovation of students' classroom behavior data collection methods, this study selected six classrooms equipped with camera equipment and

analyzed classroom teaching videos with the support of computer vision technology, which provided data support for teachers to master students' learning engagement, optimize teaching design and implement teaching intervention (Ginda et al. 2019).

According to the complexity of human behavior, it can be divided into four categories, namely posture, individual action, interactive action and group activity network. Posture is the movement of basic parts of human body, such as raising hands and standing up. This type of behavior is the least complex. Individual actions are the combination of multiple gestures, such as running, jumping, etc. Interactive actions include between people and objects, such as playing mobile phones, shaking hands, etc. Group activities refer to activities involving multiple people and objects in a scene (Rozado et al. 2017; Asadi et al. 2017), such as meeting in a conference room, marathon, etc. In the classroom scene, students' behaviors not only include basic gestures related to posture, such as raising hands, sidling, bowing down, etc. It also covers the interaction between people and objects, such as writing and playing with mobile phones.

Behavioral recognition of vision usually includes the representation of the behavior and the detection of the target. The characterization method of human joint behavior is to obtain the position information and movement information of all joints of human body through attitude estimation, and then to represent human behavior. Multi-person two-dimensional key point detection algorithm can be divided into top-down and bottom-up according to the sequence of human body detection and human key point detection (Pang et al. 2021; Teng et al. 2022). The most classic bottom-up method Open-Pose firstly detects the joints of body parts according to the maximum thermal value, and gets the human posture skeleton after connection, and puts forward the human affinity field to realize the fast connection of the joints. When the number of people in the image increases, Open-Pose algorithm (Patel et al.2022) can still maintain high efficiency and high quality to produce human posture detection results, with strong robustness.

The object detection algorithm can locate the position of the image object and give the classification result. R-CNN (Region with CNN features) (Kido et al. 2018) series algorithms combine the candidate region with convolutional neural network, and thus derive algorithms Fast R-CNN and Faster R-CNN with higher processing speed and accuracy. This kind of algorithm has the advantage of high precision, but the detection speed is slow, can not meet the real-time. Redmon et al. [7] combined the generation of candidate boxes and regression into one step, and proposed a series of representative algorithms such as YOLO v2 and YOLO v3 algorithm. This paper carried out pruning treatment on the YOLO v3 model to further reduce model parameters, improve processing speed, reduce computing resources and time consumption, and facilitate the deployment of the model while ensuring accuracy.

Jin et al. (2017) used deep convolutional neural network to analyze students' classroom expressions, and classified their emotions into sadness, happiness, neutral, anger, disgust, surprise and fear. Cheng et al. (2020) used Cohn-Kanade (CK+) facial image database to conduct deep network model pre-training, and then migrated the network according to their own application scenarios. Lei et al. (2019) proposed a multi-feature student action recognition method, which consists of local logarithmic Euclidean multivariate Gaussian (L2EMG) and Scale invariant Feature Transform (SIFT). Neupane et al. (2019) used key point information of human body and RGB (Red-Green-Blue) images to identify students' three behaviors: raising hands, standing up and sitting up. Che et al. (2022) collected real intelligent classroom environment video data, made students' class action recognition database, and conducted benchmark experiments on the database by using traditional machine learning methods and convolutional neural network. Ma et al. (2022) used C3D (Convolution 3D) network to recognize actions of students on the self-built classroom learning database. This kind of method does not make use of attitude information and interactive object information, so there are not many kinds of behavior recognition, low accuracy and slow processing speed. With the increase of the number of network layers, the deep network model is prone to over-fitting phenomenon, and the consumption of computing resources is large.

2.1 ResNet-50 Model

Theoretically, increasing the width and depth of the network can improve the performance of the network well. However, when the network reaches a certain number of layers, the classification performance will not improve with the increase of the depth of the network, but the accuracy will be reduced because of the decline of generalization performance, which is called network degradation. ResNet (residual neural network) can solve this problem, and the performance is still very good while the network is deepened (Sander et al.2021). The residual unit in the residual neural network includes Identity block and Conv block, and uses the "quick connection" in Identity to transmit information. The infrastructure of the ResNet-50 model is shown in Figure 1.

The ResNet-50 model has 50 layers, and the first convolution module is composed of 7×7 convolution layer and 3×3 maximum pooling layer. The second to fifth convolution deep modules are stacked by multiple residual learning units.

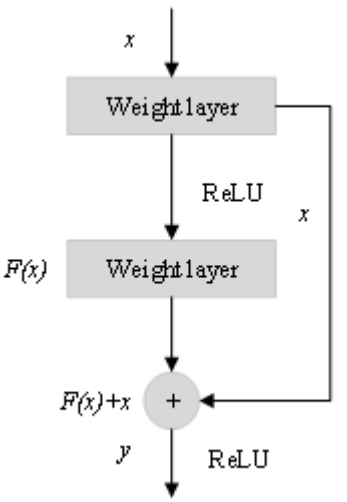
2.2 The EfficientNet-B0 Model

The EfficientNet model (Mahasin et al. 2022) is a lightweight network with good generalization performance, yielding a fixed ratio by composite scaling to balance the three dimensions of image resolution, network depth and width, while improving accuracy and efficiency. The calculation formula of dimensional composite scaling is as follows:

$$\begin{cases} depth : d = \alpha^{\phi} \\ width : w = \beta^{\phi} \\ resolution : r = \gamma^{\phi} \\ s.t. \alpha\beta^2\gamma^2 \approx 2 \\ \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{cases} \tag{1}$$

where ϕ is the composite coefficient, α, β, γ are the distribution coefficients of depth, width and resolution obtained by grid search, respectively. The optimal α, β, γ is obtained by grid search,

Figure 1. ResNet-50 network



and then the basic model is determined. Scaling the dimensions using equation (1), doubling a, doubles the number of floating-point operations. But doubling beta and y will increase the number of floating-point operations by a factor of four. The moving-flip bottleneck convolution was constructed using the compression-excitation optimization and reverse residual structure in MobileNetV2, and the EfficientNet model was established as the basic module.

3. DATA ACQUISITION AND PRE-PROCESSING

In this paper, three typical classroom behaviors of students are identified, which are sitting up, standing up and raising hands. Among them, raising hands and standing up are good signals for video recording and playback. Once the algorithm detects the existence of these two types of actions, it can switch shots. Once the action type changes back to sitting, the camera automatically switches back to the teacher's direction. At the same time, these movements also reflect the basic situation of students' classroom learning. The identification of these typical behaviors can also provide data support for the subsequent automatic teaching analysis.

A total of 90 students were selected as research objects in this study, 30 of whom were in 3 batches, and data of the same type of actions were uniformly collected. Specific collection methods are as follows: For the same batch of people, the video of a certain type of movement, such as sitting up, shall be collected first, and students shall be required to have a small range of changes, such as body tilt, head swing, etc., to ensure the diversity of the collected data (but only within the range of this type of movement). Every 5 minutes, the next movement shall be changed until the collection of the three types of movement is completed, and the next batch of students shall continue to collect. At the end of the collection, 9 videos are obtained, each of which is about 5 minutes long.

After the collection, the candidate frames of each student are obtained by the YOLO v3P target detection algorithm for the obtained videos, and then the candidate frames are cut out and manually screened, and some irrelevant and repetitive candidate frames are removed to get the data of each class of actions of each student. Finally, the data of each type of action is guaranteed to be about 130,000. At the same time, in order to ensure the reliability of experimental results, 30 additional students were selected for the collection of verification set data. The specific classroom learning behavior recognition data set D0 is shown in Table 1.

In order to ensure the effective and fast convergence of model training, the following preprocessing operations were carried out on the acquired images: the captured images with different sizes were uniformly scaled to 256×256 (blank part was filled with zero); Then, based on the center point, the scaled 256×256 images are uniformly clipped to 224×224 images. Finally, the image is de-averaged.

4. PROPOSED BEHAVIOR RECOGNITION METHOD

4.1 Key Points of Human Body

The key points of the human body can be used to represent the human skeleton and shape. The coordinate position of the key points in the image can be recognized and obtained by the algorithm,

Table 1. Classroom student behavior recognition data set D0

Gesture	Precision	Recall
Sitting	132085	4867
Standing	128274	980
Raising hand	133392	788
Average	393751	6633

so as to describe various behaviors of the human body. At present, the key points of detection are head, neck, torso, arms, legs and feet. This study adopted 18 key points of human body in COCO data format, as shown in Figure 2.

As an auxiliary means, human key point detection plays an important role in human tracking, behavior detection, gait recognition and other computer vision recognition fields. At present, the key points of human body are mainly applied in security monitoring, virtual reality, medical rehabilitation, etc., which are relatively rarely applied in the scene of classroom student behavior recognition.

At present, there are mainly three kinds of classroom student behaviors that need to be identified: sitting up, standing up and raising hands, and the posture of these three behaviors is highly distinguishable. Therefore, the key points of human body can be considered for identification.

4.2 Behavior Recognition Algorithm Based on RGB Image Classification

Classroom behavior recognition based only on key points is not easy to be interfered by background factors and has strong robustness. However, it also ignores the possible help of semantic information brought by background. Therefore, this study attempts to identify students' classroom behaviors based on RGB image classification to explore the influence and help degree of each modal data on model recognition.

The process of the RGB image classification method is similar to the data acquisition method. The real-time video stream is obtained by the YOLO v3 target detection algorithm, and then the candidate boxes are cut out and sent into the Resnet18 model which is pre-trained on ImageNet and optimized in the D0 data set for forward inference. Therefore, the model will give the probability value of each of the three behaviors that the candidate box belongs to (add up to 1), and the behavior with the largest probability value is the result obtained by the model recognition. The accuracy and recall rate of the verification set are shown in Table 2.

Figure 2. COCO data format 18 human body key points chart

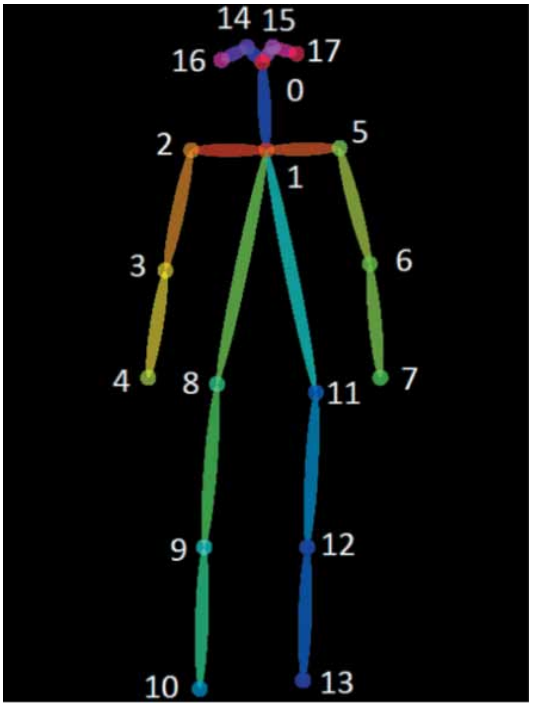


Table 2. Results of behavior recognition algorithm based on RGB image classification

Gesture	Precision	Recall
Sitting	0.9714	0.5244
Standing	0.6122	0.9550
Raising hand	0.2958	0.9368
Average	0.6236	0.8026

Among them, the accuracy of sitting and raising hands is not higher than that of the algorithm based on key points, and the average accuracy is 0.6236, which is not good. Therefore, the method of RGB image classification alone can not identify students' behaviors well.

4.3 Behavior Recognition Algorithm Based on Key Points and RGB Images

Based on single modal data such as key points or RGB images, the classroom student behavior can not be well identified. Therefore, the behavior recognition algorithm based on multi-modal data is considered, that is, the information of both key points and RGB images is combined. Because keypoint-based recognition can not be interfered by classroom background, the model can focus more on the difference between learning behaviors, while the RGB image can complete the useful background semantic information lost by the key points, which can complement each other.

The behavior recognition algorithm based on key points and RGB images mainly uses two branches to extract features respectively, then carries out feature fusion, and finally makes prediction. First, for a certain video frame extracted from the video, the candidate box of a student is obtained by YOLOv3 target detection algorithm. After pretreatment, it is divided into left and right branches. On the left is the key point branch. After the Single Person Pose Estimator (SPPE), it obtains the 54-dimensional characteristic information of this person and uses Min-Max Normalization, which aims at using linear transformation of original data. Make the result value map between 0 and 1; At the same time, the candidate frame passes through the RGB image branch on the right to extract 512 dimensional features of the last layer of the ResNet18 model except the full connection layer. After that, concat (channel connection) operation is carried out between 54-dimensional features of key points and 512-dimensional features of images, that is, direct series features, and behavior prediction is carried out based on this as the last feature.

4.4 Improved ResNet-50 Model Network

The activation function used in the traditional ResNet-50 model network is the ReLU (rectified linear unit), and the rectified linear unit can effectively prevent dispersion by taking advantage of its large gradient value during back propagation. However, ReLU also has some limitations, that is, when the input is less than or equal to 0, the output is directly 0, which is easy to make the gradient disappear. Therefore, ReLU is replaced by an expanded exponential linear unit (SeLU, scaled exponential linear unit) of the form.

$$S\left(x\right)=\lambda\begin{cases}x, & x>0\\ \theta e^{x}-\theta, & x\le 0\end{cases} \tag{2}$$

where, λ and θ are constant, $\lambda \approx 1.0507$, $\theta \approx 1.6733$. When the input is less than or equal to 0, the output is no longer directly 0, which not only has the advantage of unilateral inhibition of ReLU activation function, but also prevents the gradient from disappearing.

The original image with resolution of 224×224 is input into ResNet50 model for feature extraction, and a feature map of 7×7×2048 is obtained. In order to extract more sufficient image features, the obtained feature map is used as the input of adding parallel convolution module, and the convolution operation is continued.

The added parallel convolution module is an effective local topological network, which performs multiple convolution operations and pooling operations in parallel on the output feature graphs of the ResNet-50 model.

Convolution kernels of different sizes can obtain different features of information from different perceptual domains in the input image. Therefore, in order to effectively extract features, convolution kernels of 1×1, 3×3 and 5×5 sizes are used to construct three parallel convolution channels. Set all the convolutional steps as 1 and fill in the completion form, which will not change the feature dimension. Then, the output of three parallel channels is spliced to obtain the final feature map. Finally, the final predicted output is obtained through the average pooling layer and softmax layer. The improved parallel convolutional module structure is shown in Figure 3.

In recent years, attention mechanism model has been widely used in deep learning, especially in image processing. SeNet module, which can be called SE module, is a lightweight module whose goal is to improve the presentation quality generated by the network, including compression, firing and scaling operations. Its structure is shown in Figure 4. The module mainly compressed the original

Figure 3. Parallel convolutional module structure

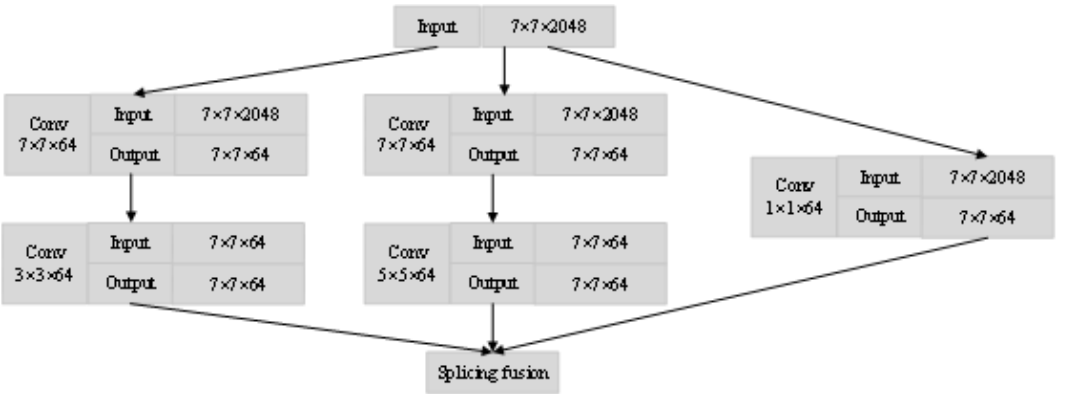
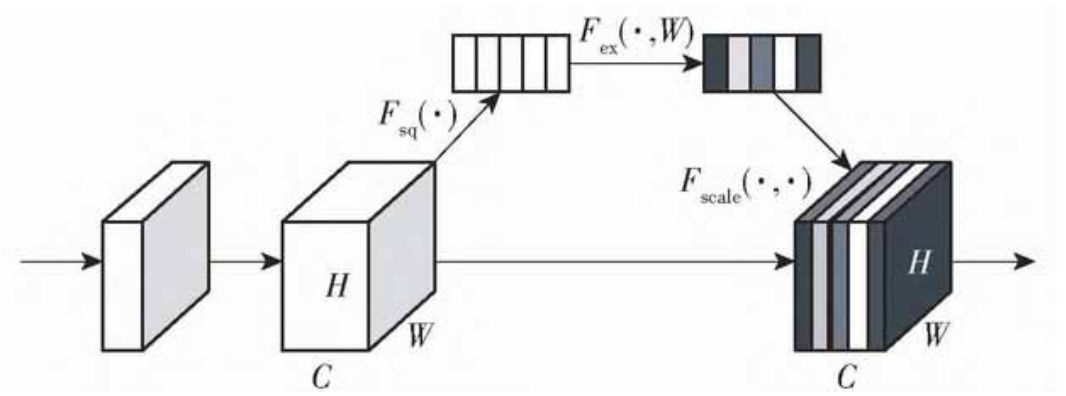


Figure 4. SeNet model



H×W×C (H is the height, W is the width, C is the number of channels) dimension feature map into 1×1×C, equivalent to compress H×W into one dimension, so as to obtain the global vision of H×W, wider feeling area. The weights are then generated for each feature channel by firing operations.

SeNet module and residual module are combined to replace the standard residual in the original network. On the basis of adding the existing network without disrupting the original main structure of the network, more attention is paid to the image features. The residual module added to the SeNet module is shown in Figure 5.

4.5 Bilinear Fusion Network

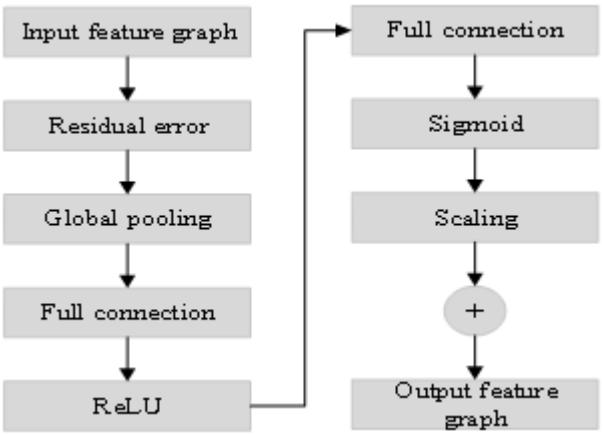
In the study on the identification of students' distracted behavior in class, only 30 students were used in the data set. Each student has multiple pictures of different behavior categories, and the movement changes slightly among different behaviors, and the difference between the categories of different behaviors is not obvious. For this kind of fine-grained image classification, if we want to improve the recognition accuracy, the difficulty will be greatly increased. Among them, bilinear convolutional neural network (BCNN, bilinear CNN) (Lin et al.2017) model is a very effective method to improve the accuracy, and different models can extract different features. Therefore, a bilinear fusion network model based on the improved ResNet-50 model and the improved EfficientNet-B0 model is proposed, and two parallel CNN single models are used as feature extractors respectively for feature extraction. This model provides stronger feature representation than the single model, and can be optimized end-to-end to achieve higher performance than the single model.

The used BCNN model is:

$$B = \left(\omega_r, \omega_e, \rho, S\right) \tag{3}$$

where ω_r and ω_e are feature extraction functions, indicating that image I and a position l on the image are mapped to a feature of C×D dimension. S is the classification function. ρ is the pooling function. The same image is entered into the improved ResNet-50 model and the improved EfficientNet-B0 model respectively, and the feature graph output is obtained at the last activation operation of the network. The model performs matrix cross product operations for the corresponding position on the feature graph and then obtains bilinear features for that position. The formula for calculating bilinear characteristics is:

Figure 5. Residual module with the SE module



$$b\left(l,I,\omega_r,\omega_e\right)=\omega_r\left(l,I\right)^T\ \omega_e\left(l,I\right) \tag{4}$$

Summing the features at different locations will eventually give a bilinear feature descriptor f of global L, i.e.:

$$f\left(I\right)=\sum_{l\in L}b\left(l,I,\omega_r,\omega_e\right) \tag{5}$$

Finally, softmax classification function is used to classify decision among feature descriptors. As shown in Figure 6, the BCNN model can be used for end-to-end training, and one picture has been improved respectively. The two independent models ResNet-50 and EfficientNet-B0 obtain two different feature descriptions and perform bilinear fusion, enabling the network to achieve higher feature extraction capabilities, thus achieving better classification results.

5. EXPERIMENTS AND ANALYSIS

In the previous model training, only sitting, standing up and raising hands were used. But the actual situation will be more complicated, may be sitting this behavior also includes lying on the table, reading, writing these behaviors. So consider folding table-leaning, reading, and writing into the sitting category. Data screening was performed manually to remove some duplicate data and obtain a new dataset D1, as shown in Table 3.

The experimental environment is Inter (R) Core (TM) i5-3317U-CPU, 64-bit,Windows 10 system. This study is a classification problem, and the most common evaluation indexes include accuracy, recall rate and confusion matrix. In the actual classification process, there are several cases: true example, false negative example, false positive example and true negative example.

In neural networks, optimizers are usually used to optimize the model. Stochastic gradient descent (SGD) algorithm, adaptive gradient algorithm (AdaGrad), and adaptive moment estimation (Adam)

Figure 6. Bilinear convolutional neural network model

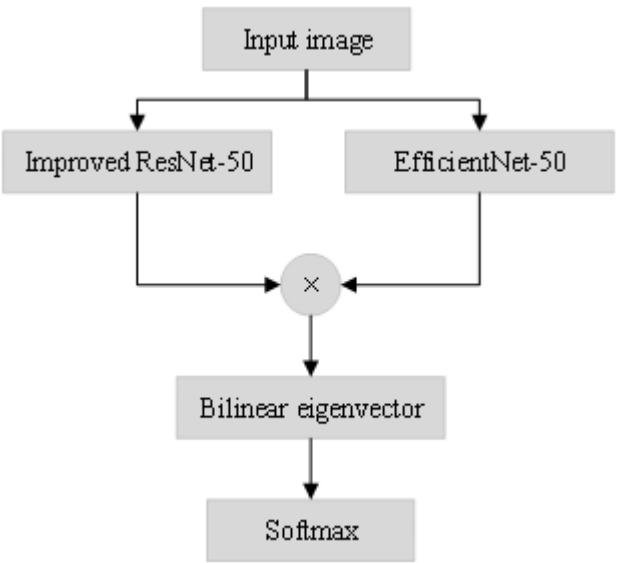


Table 3. Student behavior recognition data set D1

Gesture	Training	Verification
Sitting	236934	4866
Standing	43379	980
Raising hand	16936	788
Total	297249	6634

algorithm are used for training tests of the ResNet-50 model. It can be seen from Table 4 that the Adam optimizer and SeLU activation function are the optimal choices.

The training sets are entered into the ResNet-50 model, EfficientNet-B0, the improved ResNet-50 model and improved ResNet-50+EfficientNet-B0 model for training. Table 5 shows the accuracy of the training using the four models.

It can be seen that the improved ResNet-50+EfficientNet-B0 fusion model shows the highest accuracy, with an average accuracy of 97% after 20 training iterations. The average results of multiple experiments show that the accuracy of the modified ResNet-50 model can be improved by 5%. Table 6 shows the results of testing the four models.

To verify the performance of the network model, VGG16, Inception-v3 single model and BP-CNN, Inception-V3-Resnet152-v2 fusion model network were used for control experiments under the same data set. Table 7 shows the overall classification accuracy obtained after testing eight network models under the ten-fold crossover experiment.

The above experimental results show that the improved ResNet50+EfficientNet-B0 fusion model achieves the highest overall recall rate and overall accuracy, and has relatively high recall rates for each type of student behavior.

Table 4. Classification accuracy under different activation functions and optimizers/%

Model	Activation Function	Optimizer	Overall Accuracy
ResNet-50	ReLU	SGD	87.6
ResNet-50	ReLU	AdaGrad	88.2
ResNet-50	ReLU	Adam	88.6
ResNet-50	SeLU	SGD	89.3
ResNet-50	SeLU	AdaGrad	89.7
ResNet-50	SeLU	Adam	90.4
New ResNet-50	SeLU	Adam	95.3

Table 5. Accuracy of model training/%

Model	EfficientNet-B0	ResNet-50	Improved ResNet-50	ResNet-50+EfficientNet-B0
2	19.6	29.8	28.7	26.5
10	80.4	81.7	86.1	87.2
14	81.9	82.8	90.7	91.4
20	88.7	91.3	95.5	97.0

Table 6. Recall rate under different models/%

Model	Sitting	Standing	Raising Hand
EfficientNet-B0	92.8	93.0	92.5
ResNet-50	83.5	82.3	86.7
Improved ResNet-50	88.5	93.5	93.6
ResNet-50+EfficientNet-B0	94.2	96.4	97.6

Table 7. Accuracy under different models/%

Model	Accuracy
VGG16	64.3
Inception-v3	73.2
ResNet-v3	90.4
Improved ResNet-50	95.3
EfficientNet-B0	94.9
BP-CNN	85.4
Inception-V3-Resnet152-v2	92.6
ResNet-50+EfficientNet-B0	97.8

6. CONCLUSION

Using behavior recognition algorithm to identify classroom student behavior can help teachers improve teaching methods and promote the application of intelligent video recording and broadcasting system, which is of great significance. This paper studies the influence of multi-modal data on improving the accuracy and recall rate of classroom students' behavior recognition through multiple sets of experiments, and finds that the behavior recognition method based on multi-modal data, such as human key points, RGB images, etc., has a certain improvement in effect. On this basis, through effective post-processing of the data, such as combining multiple types of data, re-labeling the data with large losses according to the ranking of training loss values and setting weights between different categories, the effect of the recognition task can be further improved, and the effect is significantly improved. The class student behavior recognition algorithm based on multi-modal data in this paper can be widely used to solve common problems such as recognition and classification in practical work, and has certain practical significance.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

Ahmad, M., & Lee, S. W. (2008). Human action recognition using shape and CLG-motion flow from multi-view image sequences. *Pattern Recognition*, *41*(7), 2237–2252. doi:10.1016/j.patcog.2007.12.008

Asadi-Aghbolaghi, M., Clapes, A., & Bellantonio, M. (2017). A survey on deep learning based approaches for action and gesture recognition in image sequences. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017).* IEEE.

Che, B., Li, X., Sun, Y., Yang, F., Liu, P., & Lu, W. (2022). A database of students' spontaneous actions in the real classroom environment. *Computers & Electrical Engineering*, *101*, 108075. doi:10.1016/j.compeleceng.2022.108075

Cheng, S., & Zhou, G. (2020). Facial expression recognition method based on improved VGG convolutional neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, *34*(07), 2056003. doi:10.1142/S0218001420560030

Cheng, Y. (2022). Video-based Student Classroom Classroom Behavior State Analysis. *International Journal of Education and Humanities*, *5*(2), 229–233. doi:10.54097/ijeh.v5i2.2146

Ginda, M., Richey, M. C., Cousino, M., & Börner, K. (2019). Visualizing learner engagement, performance, and trajectories to evaluate and optimize online course design. *PLoS One*, *14*(5), e0215964. doi:10.1371/journal.pone.0215964 PMID:31059546

Groccia, J. E. (2018). What is student engagement? *New Directions for Teaching and Learning*, *2018*(154), 11–20. doi:10.1002/tl.20287

Jin, K. H., McCann, M. T., Froustey, E., & Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, *26*(9), 4509–4522. doi:10.1109/TIP.2017.2713099 PMID:28641250

Jisi, A., & Yin, S. (2021). A new feature fusion network for student behavior recognition in education. *Journal of Applied Science and Engineering*, *24*(2), 133–140.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188. doi:10.1017/S0140525X10003134 PMID:21864419

Kido, S., Hirano, Y., & Hashimoto, N. (2018). Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN). In *2018 International workshop on advanced image technology (IWAIT).* IEEE.

Lei, F., Wei, Y., & Hu, J. (2019). Student action recognition based on multiple features. In *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData).* IEEE.

Lin, T. Y., RoyChowdhury, A., & Maji, S. (2017). RoyChowdhury A, Maji S. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1309–1322. doi:10.1109/TPAMI.2017.2723400 PMID:28692962

Luo, H., Wu, C., & He, Q. (2015). Research on cultivating medical students' self-learning ability using teaching system integrated with learning analysis technology [J]. *International Journal of Clinical and Experimental Medicine*, *8*(8), 14542. PMID:26550446

Ma, Y., Wei, Y., Shi, Y., Li, X., Tian, Y., & Zhao, Z. (2022). Online Learning Engagement Recognition Using Bidirectional Long-Term Recurrent Convolutional Networks [J]. *Sustainability (Basel)*, *15*(1), 198. doi:10.3390/su15010198

Mahasin, M., & Dewi, I. A. (2022). Comparison of CSPDarkNet53, CSPResNeXt-50, and EfficientNet-B0 Backbones on YOLO V4 as Object Detector[J]. International Journal of Engineering. *Science and Information Technology*, *2*(3), 64–72.

Meng, X., Wang, X., & Yin, S. (2023). Few-shot image classification algorithm based on attention mechanism and weight fusion. *Journal of Engineering and Applied Sciences (Asian Research Publishing Network)*, *70*(1), 1–14.

Neupane, B., Horanont, T., & Hung, N. D. (2019). Deep learning based banana plant detection and counting using high-resolution red-green-blue (RGB) images collected from unmanned aerial vehicle (UAV). *PLoS One*, *14*(10), e0223906. doi:10.1371/journal.pone.0223906 PMID:31622450

Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., & Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews. Genetics*, *24*(2), 125–137. doi:10.1038/s41576-022-00532-2 PMID:36192604

Pang, Y., Wang, Q., & Zhang, C. (2021). Time-frequency domain pattern analysis of Tai Chi 12 GONG FA based on skeleton key points detection. In *2021 International Conference on Neural Networks, Information and Communication Engineering*. SPIE.

Patel, M. C., & Kalani, N. B. (2022). Personal Exercise Assistant: Correcting Exercise Posture using Modified Open pose. *Mathematical Statistician and Engineering Applications*, *71*(4), 10347–10358. doi:10.17762/msea.v71i4.1871

Rozado, D., Niu, J., & Lochner, M. (2017). Fast human-computer interaction by combining gaze pointing and face gestures. *ACM Transactions on Accessible Computing, 10*(3), 1-18.

Sander, M. E., Ablin, P., & Blondel, M. (2021). Momentum residual neural networks. In *International Conference on Machine Learning*. PMLR.

Sander, M. E., Ablin, P., & Blondel, M. (2021). Momentum residual neural networks. In *International Conference on Machine Learning*. PMLR.

Wang, L., Shoulin, Y., & Alyami, H. (2022). *A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images*. 10.1002/gdj3.162

Wu, J., & Jiang, X. (2020). A study on Strategies of Stimulating Students' Participation in English Teaching Based on Computer Information Technology. *Journal of Physics: Conference Series*, *1578*(1), 012079.