Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25725

Classroom Attention Detection Using YOlOv8 and ResNet-based Facial Expression Recognition Model

Chenyu Xing

Hangzhou Foreign Languages School, Hangzhou, China
himjavascript@outlook.com

Abstract. This paper proposes a classroom attention detection model based on YOLOv8 and ResNet, aimed at evaluating students' engagement during class by analyzing facial expressions. The model consists of two main components: the YOLOv8 and ResNet classification models. YOLOv8 is used to localize facial regions in real time, and ResNet is used to recognize students' emotional states. Emotional states are categorized into seven facial expressions: "Surprise", "Fear", "Disgust", "Happy", "Sad", "Angry", and "Neutral". Positive expressions such as "Happy", "Surprise", and "Neutral" are classified as "Good" attention states, while negative expressions are classified as "Bad" attention states. Experimental results show that the proposed model achieves high precision and recall, with a precision of 0.88649, recall of 0.6654, mAP@50 of 0.7567, and mAP@50-95 of 0.41387. These results indicate that the model can effectively assess students' classroom attention, providing valuable feedback for teachers to monitor student engagement. In conclusion, this study demonstrates the potential of using facial expression recognition for classroom attention detection. Future work will focus on improving the model's real-time performance, adaptability across different environments, and integrating multimodal data for more accurate and personalized attention evaluation.

Keywords: yolov8, resnet, Expression Recognition, Classroom Attention

1. Introduction

This document is Classroom attention detection refers to assessing students' attention status in class by analyzing students' facial expressions, behaviors and other data. With the development of education informatization, how to efficiently monitor students' classroom participation has become a hot issue in the field of education. While traditional classroom observation often relies on teachers' subjective judgment, facial expression analysis technology can objectively and real-time assess students' attention status. In this study, we used facial expression detection combined with classification modeling to achieve classroom attention detection, which can accurately assess students' participation without teacher intervention. However, facial expressions are affected by a variety of factors, such as individual differences and blurred expressions, which makes classroom attention detection very difficult.

To solve the problem of classroom attention detection, existing methods are mainly categorized into the following 3 classes.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ25725

First, traditional methods are based on computer vision. These methods extract facial features (e.g., Haar features, LBP, etc.) by hand and use traditional machine learning models for classification. The method is simple to implement but has poor robustness in complex environments. Related literature proposed a facial expression recognition method based on Haar features, which showed good accuracy but was more sensitive to light and expression changes [1-2].

Secondly, facial expression detection methods based on convolutional neural networks (CNN) and other deep learning techniques have made significant progress. This method can better adapt to complex environments by automatically learning features without relying on manually designed features. Elbawab and Henriques used deep convolutional networks to extract facial expression features and achieved more excellent classification results [3].

The third is multimodal approaches method, some studies have also combined multiple information such as speech and behavior to improve detection accuracy [4]. Such approaches tend to provide a more comprehensive attention assessment by fusing data from different modalities. Relevant literature proposed a multimodal emotion recognition system, which improves recognition accuracy by combining facial expressions and speech features [5].

Although the existing methods have solved the problem of classroom attention detection to a certain extent, there are still some shortcomings. First, the feature extraction process of traditional methods relies on manual design, lacks flexibility and adaptability, and performs unstably under complex lighting and angle changes. Second, although deep learning methods can automatically extract features, they have high data requirements, long training time, and high hardware requirements for models. In addition, most of the existing methods focus on the information of a single modality and neglect the fusion of multimodal data, resulting in insufficient accuracy and robustness.

In order to solve the above problems, this study proposes a combined model based on YOLOv8 and ResNet, which utilizes YOLOv8 for efficient target detection, firstly locates the student's facial region, and then performs expression classification through ResNet network to determine the student's attention state. By combining the real-time detection capability of YOLOv8 and the deep feature learning capability of ResNet, we are able to recognize students' expressions stably and accurately under various environmental conditions and accurately assess their listening status.

This thesis proposed a novel classroom attention detection model is proposed, which effectively solves the robustness problem of traditional methods in complex environments through the combination of YOLOv8 and ResNet. In addition, a classification model based on 7 types of expressions clarifies the differentiation criteria between positive and negative listening states, providing a new technical means for classroom behavior analysis. Moreover, the superiority of the method on different datasets is verified through a large number of experiments, and feasibility support is provided for the application in real classrooms.

2. Related work

This chapter will briefly introduce the key techniques related to classroom attention detection modeling, including YOLOv8, ResNet, and techniques related to expression recognition.

2.1. YOLOv8

YOLO (You Only Look Once) is a popular target detection algorithm, which is characterized by the fact that target localization and classification can be accomplished through a single forward propagation, thus realizing efficient and real-time target detection. YOLOv8 is the newest version of

DOI: 10.54254/2755-2721/2025.BJ25725

the YOLO series, which is optimized in terms of the accuracy and speed of target detection. It is capable of better target detection in complex environments through improved neural network structure and more efficient inference algorithms. YOLOv8 is especially suitable for real-time detection tasks, and is able to quickly localize the target region, thus providing accurate facial region data for subsequent expression recognition [6-7]. In this project, YOLOv8 is used to detect the facial regions of students in the classroom in real time to provide input for the expression classification model.

2.2. ResNet

ResNet (Residual Network) is a deep convolutional neural network, which adopts Residual Learning (RL) to effectively solve the problem of gradient vanishing during deep network training. By introducing the residual module, ResNet allows the network to be trained at a deeper level, significantly improving the performance of the model. ResNet performs well in image classification and feature extraction tasks, and is able to automatically extract high-level features in facial expressions. In this study, ResNet is used as part of the expression classification model to analyze the facial region images detected by YOLOv8 for classification of emotional states.

2.3. Expression recognition techniques

Expression recognition is an important task in affective computing, aiming at recognizing human emotional states by analyzing their facial expressions. Current expression recognition methods mainly rely on deep learning techniques, especially Convolutional Neural Networks (CNN), which can automatically extract features from images and recognize complex facial changes. Expression recognition is usually divided into two steps. First, facial regions are localized using target detection algorithms (e.g., YOLOv8). Then, features of facial expressions are analyzed by classification models (e.g., ResNet) to determine the individual's emotional state. Commonly categorized facial expressions include “Happy”, “Sad”, “Angry”, “Neutral”, and so on. Expression recognition has been widely used in the field of education to effectively assess students' classroom attention and mood swings by monitoring their emotional changes in real time [6-10].

3. The proposed method

This chapter will briefly describe the design and implementation process of the classroom attention detection model based on YOLOv8 and ResNet proposed in this paper.

3.1. Model design

The model proposed in this paper is mainly composed of two parts: target detection module and expression classification module. First, facial detection is performed using YOLOv8. YOLOv8 is able to detect students' facial regions in real-time in complex classroom environments and outputs detection frames to precisely locate each student's facial image. Next, these facial images are fed into a ResNet model for expression classification. ResNet extracts facial features through a deep convolutional network and classifies them into seven expressions: Surprise, Fear, Disgust, Happy, Sad, Angry, and Neutral.

Based on the classification results of these expressions, the model further determines the attention state of the students: if the expression is positive (e.g., Happy, Surprise, Neutral), it is determined as “Good” listening state; if the expression is negative (e.g., Sad, Angry, Fear, Disgust), it is

DOI: 10.54254/2755-2721/2025.BJ25725

determined as “Bad” listening state. If the expression is negative (e.g. Sad, Angry, Fear, Disgust), it will be recognized as “Bad” listening state [8-9].

3.2. Model implementation

The model implementation is divided into the following steps:

Data preprocessing: First, collect and label the image dataset containing students' facial expressions. The images are normalized and enhanced to improve the robustness of the model.

Target detection: target detection is performed on the input images using YOLOv8 to accurately identify and localize the student's facial regions. YOLOv8 is pre-trained to perform the detection task efficiently and accurately.

Expression Classification: The facial regions detected by YOLOv8 are fed into ResNet for classification, which extracts facial expression features through its deep network architecture and classifies them into seven types of expressions.

State Judgment: Based on the result of expression classification, set rules to judge the attention state of students, and finally output “Good” or “Bad” listening state.

3.3. Model optimization

In order to improve the accuracy and real-time performance of the model, we designed an efficient data flow path between YOLOv8 and ResNet and optimized the network structure. The parameters of YOLOv8 were adjusted to balance the detection accuracy and speed to ensure the real-time performance of the model in the real classroom environment. Meanwhile, the training process of ResNet was optimized to enhance its ability to learn different expression features.

4. Results

This chapter shows the experimental results of the classroom attention detection model based on YOLOv8 and ResNet and analyzes its performance.

In the experiments, we used a standard dataset containing students' facial expressions, YOLOv8 for target detection and ResNet for expression classification. During model training, cross-validation was used to ensure the stability of the results, and the four metrics of precision, recall, mAP@50 and mAP@50-95 were evaluated. The experimental results are shown in Table 1 and Fig. 1.

Table 1. The result of Value for each metric

Metrics	Value
Precision (B)	0.88649
Recall (B)	0.6654
mAP@50 (B)	0.7567
mAP@50-95 (B)	0.41387

DOI: 10.54254/2755-2721/2025.BJ25725

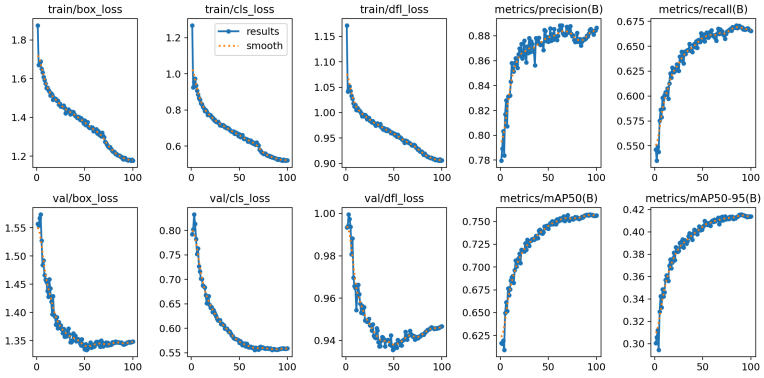

Figure 1. Evaluation results for precision, recall, mAP@50 and mAP@50-95

Table 1 shows that Precision is 0.88649, indicating that the model is accurate in predicting "Good" listening status. The Recall is 0.6654, which means that the model is able to recognize the students who are actually in the "Good" listening status, but there are still some omissions. mAP@50 is 0.7567, indicates that the model has found a better balance between detection accuracy and recall. mAP@50-95 is 0.41387, indicates that the model's performance is stable under different IoU thresholds, which is suitable for practical application scenarios.

The experimental results show that the classroom attention detection model based on YOLOv8 and ResNet proposed in this paper performs well in the task of target detection and expression classification, and is able to effectively assess students' listening status. Although there is room for improvement in the recall rate, the overall performance has reached the expected effect and has high practical application value.

5. Conclusion

In this paper, a classroom attention detection model based on YOLOv8 and ResNet is proposed to effectively assess students' classroom participation status through real-time facial expression analysis. The experimental results show that the model performs well in target detection and expression classification tasks, with high levels of precision and recall, and can provide powerful technical support for classroom management.

For the future research direction, in multimodal fusion, students' language, behavior and other information are combined to further improve the precision and robustness of the model. In real-time optimization, optimize the model inference speed to make it better adapt to large-scale classroom application scenarios. For cross-environment adaptation, enhance the model's performance under different lighting and angles to improve its applicability in real classrooms. In addition, the attention assessment model is customized according to individual student differences to provide more accurate feedback on classroom behaviors. Through these optimizations, the model is expected to be applied to a wider range of educational environments, providing more support for improving teaching quality.

References

[1] Zhu, H., Zhao, J., & Niu, L. (2022, December). An efficient model for student behavior recognition in classroom. In 2022 International Conference on Intelligent Education and Intelligent Research (IEIR) (pp. 142-147). IEEE.

[2] Betto, I., Hatano, R., & Nishiyama, H. (2023). Distraction detection of lectures in e-learning using machine learning based on human facial features and postural information. Artificial Life and Robotics, 28(1), 166-174.

DOI: 10.54254/2755-2721/2025.BJ25725

[3] Elbawab, M., & Henriques, R. (2023). Machine Learning applied to student attentiveness detection: Using emotional and non-emotional measures. Education and information technologies, 28(12), 15717-15737.
[4] Duan, Y., & Zhang, S. (2024, July). Student classroom behavior recognition based on bidirectional feature fusion. In Third International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2024) (Vol. 13181, pp. 116-121). SPIE.
[5] Liu, D., Wang, Z., Wang, L., & Chen, L. (2021). Multi-modal fusion emotion recognition method of speech expression based on deep learning. Frontiers in Neurorobotics, 15, 697634.
[6] Sun, J., Li, S., & Zhang, J. (2024). Classroom Behavior Recognition and Research Based on DLKAS-YOLO8n. Academic Journal of Computing & Information Science, 7(11).
[7] Liu, Q., Jiang, R., Xu, Q., Wang, D., Sang, Z., Jiang, X., & Wu, L. (2024). Yolov8n_bt: Research on classroom learning behavior recognition algorithm based on improved yolov8n. IEEE Access.
[8] Wang, L., Chen, L., & Liu, W. (2024, September). A Study on Student Behavior Recognition Based on YOLOv7. In 2024 2nd International Conference on Signal Processing and Intelligent Computing (SPIC) (pp. 1158-1163). IEEE.
[9] Fan, X., & Fan, B. (2024, March). STUEFF-YOLOv5s: A Lightweight Method for Deep Identification of Students' Behavior in the Classroom. In Proceedings of the 2024 3rd International Conference on Cyber Security, Artificial Intelligence and Digital Economy (pp. 401-405).
[10] Wang, Z., Yao, J., Zeng, C., Wu, W., Xu, H., & Yang, Y. (2023). Learning behavior recognition in smart classroom with multiple students based on YOLOv5. arXiv preprint arXiv: 2303.10916.