Towards Analyzing Student Engagement: A Deep Learning System for Classroom Behavior Recognition

Linying Yan

Xi'an Peihua University, Xi'an 710125, Shaanxi, China

Abstract: ***This study proposes a deep-learning-based student classroom behavior recognition system aimed at accurately detecting and analyzing student classroom behaviors through intelligent means. The system adopts a modular design covering five modules: data acquisition, preprocessing, model training, behavior recognition, and visual interaction. Centered on the YOLOv8 algorithm, model performance is optimized by adjusting parameters such as iteration count and learning rate. A dataset of 3200 training images and 800 test images is split at an 8:2 ratio. Techniques including Mosaic data augmentation, adaptive anchor calculation, and SPP/PAN feature fusion are employed to enhance the model's multi-scale object detection capability. The loss function comprises bounding-box loss, class loss, and confidence loss to ensure detection accuracy.***

Keywords: Deep learning; Student classroom behavior; Recognition system.

1. INTRODUCTION

Under the leap-forward development trend of information technology, the education sector is gradually moving toward digitalization and intelligence, and the traditional classroom teaching model can no longer meet the demands of current educational reforms. How to accurately grasp students' classroom learning focus and precisely assess their learning status has become a key issue that teachers urgently need to solve. Gao and Gorinevsky (2020) developed probabilistic modeling for optimizing resource mix with variable generation and storage [1]. Robotics research has seen substantial progress through Guo's (2025) work on optimal trajectory control using deterministic AI for robotic manipulators [2], real-time data completion for motion recognition with LSTM [3], and robot-environment interaction modeling with Tao [4]. Healthcare transformation is evidenced by Wei et al. (2025), who implemented AI-driven intelligent health management systems in telemedicine [5]. Network infrastructure benefits from Zhang et al.'s (2025) MamNet, a novel hybrid model for time-series forecasting and frequency pattern analysis in network traffic [6]. Computer vision has advanced through Peng and Chen's (2024) dual-augmentor framework for domain generalization in 3D human pose estimation [7] and their earlier work on RAIN for black-box domain adaptation [8]. Object recognition systems were enhanced by Chen et al. (2022) with one-stage object referring incorporating gaze estimation [9]. Business applications include Zhang, Jingbo et al.'s (2025) AI-driven sales forecasting for gaming industry advertising markets [10] and Zhang, Yuhan's (2025) CrossPlatformStack for enabling high-availability deployment across meta services [11]. Advertising technology has evolved with Hu's (2025) AdPercept for visual saliency and attention modeling in 3D ad design [12], while privacy preservation is addressed by Li, Lin, and Zhang (2025) through federated learning and differential privacy frameworks for advertising personalization [13]. Network testing automation is achieved by Tu (2025) with AutoNetTest for intelligent 5G network diagnosis [14], and platform stability is ensured by Zhu's (2025) ReliBridge using scalable LLM backbones for small businesses [15]. Content creation is advanced by Hu (2025) through few-shot neural editors for 3D animation in SMEs [16]. Industrial applications feature Tan et al.'s (2024) damage detection system using deep transfer learning and ensemble classifiers [17], while manufacturing optimization is enhanced by Xie and Chen's (2025) Maestro, a multi-agent system for task recognition in production lines [18].

2. SYSTEM FUNCTIONAL MODULE DESIGN

This system design is based on modular design; the framework covers a data acquisition module, a data preprocessing module, a model training module, a behavior recognition module, and a visualization interaction module.

2.1 Data Acquisition Module

The data acquisition module collects images of student classroom performance from multiple information sources. While grabbing open image data from open-source projects such as GitHub, it also captures classroom images from various school surveillance cameras to ensure the richness and objectivity of the collected sample data. The system plans to collect classroom behaviors such as "raising hand," "lowering head," "turning around," and "standing up."

2.2 Data Preprocessing Module

The data preprocessing module mainly performs a series of processing steps on the acquired data to improve data timeliness. It cleans the collected data, removes useless images with poor clarity or completeness, and then uses data augmentation methods—such as brightness transformation, blur transformation, and image rotation—to enhance model robustness. Subsequently, the LabelImg image annotation method is employed to provide detailed labels for classroom behaviors like "raising hand," "lowering head," "turning around," and "standing up," supplying the model training with sufficient high-quality annotated information.

2.3 Model Training Module

The model training subroutine adopts the YOLOv8 behavior detection model as its algorithm. During training, model parameters such as iteration count and learning rate are adjusted to optimize the model and improve detection accuracy.

2.4 Behavior Recognition Module

The behavior recognition module integrates the trained YOLOv8 model, enabling real-time behavior detection on input images and videos, accurately identifying student classroom behavior types, and counting the frequency of each behavior.

2.5 Visualization Interaction Module

The visualization interaction module, designed with PyQt5+Gradio, provides teachers with a visual and interactive human-machine interface for detection results. Teachers can upload images or videos or use a camera for detection within the interface. The system directly displays the detection results visually, including behavior categories, confidence scores, location information, and statistical data, and can export the results to generate reports [3].

3. STUDENT CLASSROOM BEHAVIOR RECOGNITION MODEL DESIGN

3.1 Data Acquisition and Preprocessing

As a prerequisite for model training, the annotated training set is prepared. In this study, classroom behaviors such as "raising hand," "lowering head," "turning around," and "standing up" are defined as follows:

Figure 1: Classroom Behavior Diagram

Use the annotation tool LabelImg to annotate each of the 4,000 augmented images one by one. Details are as follows:

Figure 2: Annotation Diagram

The annotation process strictly follows the annotation rules to prevent any deviation when labeling the target behavior bounding boxes. After annotation, a corresponding annotation txt file is generated, containing the target behavior name, the center coordinates of the bounding box, the aspect ratio of the bounding box, and other data. A class.txt file is also created to record all behavior names, providing the correct class correspondence for the next step of model training.

3.2 Model Design and Training

The system experimental environment for this test, including the operating system, CPU, programming language, framework, and IDE:

Table 1: Experimental Environment Information Table

项目	配置
操作系统	Windows 10 专业版
CPU	Intel(R)Core(TM)i5.10300H CPU @ 2.50GHZ
编程语言	Python 3.8
框架	Pytorch 1.2.1
IDE	Pycharm 2023.3.4

3.2.1 YOLOv8 Network Model

The classroom behavior recognition method adopted by this system centers on an improved YOLOv8 algorithm, integrating multiple advanced technologies to achieve efficient and accurate identification of student classroom behaviors.

For input data, the Mosaic augmentation method is employed: four random images are stitched together to form a new training sample, increasing both the sample count and the difficulty, so the trained model gains better stability and robustness. Adaptive anchor calculation and adaptive image scaling are also added, enabling the model to handle detection tasks of varying scales and behaviors, making detection more accurate. The backbone consists of a Focus structure and a CSP structure. The Focus structure downsamples the input image via strided sampling while preserving fine details; the CSP structure splits the feature map—one part undergoes convolution while the other is passed through unchanged—then recombines the processed and unprocessed maps, reducing computation and improving feature-extraction efficiency [4]. The neck uses SPP and PAN. SPP performs multi-scale pooling on the feature maps for multi-scale fusion, enhancing the model's ability to detect objects of different sizes; PAN is an improved version of FPN that fuses features from different network layers to further boost detection accuracy. The YOLOv8 loss function comprises three main components: bounding-box loss, classification loss, and confidence loss.

(1) Bounding-box loss: This loss measures the positional deviation between the predicted bounding box and the ground-truth bounding box. It ensures the model accurately predicts the location of the target object, as shown in Equation 1:

$$CIOU_{1,0:S}1 - CIOU1\left(10U - \frac{Distance2^2}{DistanceC^2} - \frac{v^2}{(1-10U)+v}\right) \quad (1)$$

In object detection tasks in computer vision, evaluating how well a predicted bounding box matches the actual one

typically involves the following metrics:

C: This refers to the smallest enclosing rectangle formed around the predicted and ground-truth bounding boxes. Distance_2: This represents the straight-line Euclidean distance between the centers of the predicted and ground-truth bounding boxes. Distance_C: This denotes the length of the diagonal of rectangle C.V: This parameter evaluates the consistency of width-to-height ratios. IOU: Intersection over Union, the ratio of the overlapping area between the predicted and ground-truth bounding boxes to the union of their total areas.

(2) Classification loss: Indicates the accuracy of the object class within the box. By default, the cross-entropy function is used for calculation, as shown in Equation 2.

$$p\begin{cases} p, y-1 \\ 1-p, y=0 \end{cases} \tag{2}$$

Equation 2 can be simplified to Equation 3:

$$L = -\log p \tag{3}$$

In the above Sections 4.2 and 4.3, variable y represents the label of the input sample, where positive samples are marked as 1 and negative samples as 0. Variable p denotes the model's predicted probability that the input sample is positive.

(3) Confidence loss: Reflects the probability of an object existing in the predicted box, ranging from 0 to 1. A higher value indicates a greater likelihood of detecting a target in the corresponding predicted box.

Through the synergistic application of the above technologies, the system achieves precise recognition of student classroom behaviors, providing strong technical support for classroom management and teaching quality improvement [5].

3.2.2 Model Training

During training, the previously organized dataset is split into training and test sets according to 8:2, with 3,200 images for training and 800 for testing, stored in VOC-format folders for reading and processing.

Additionally, adjust the training parameters in the main function as needed, such as iteration count and image size. Through multiple training runs and parameter tuning, ensure the model maintains high recognition accuracy on the training set. Monitor changes in loss functions like bounding-box loss, classification loss, and confidence loss during training, dynamically adjusting model parameters to guarantee convergence and generalization. The final trained model demonstrates strong recognition performance on the test set, providing a basis for accurate classroom behavior identification.

4. EXPERIMENTAL ANALYSIS

This experiment evaluates the overall performance of the deep-learning-based student classroom behavior recognition system. The detailed analysis of system performance mainly covers the following aspects:

4.1 Recognition Accuracy Analysis

In addition to the 94.9% average accuracy on the test set, on-site tests were conducted in real classroom settings. Multiple lessons from different classes and courses were selected for real-time behavior assessment, and the system's accuracy reached over 93%. Most behaviors of most students were correctly identified; for example, in a 50-student classroom, 47 were correctly recognized by the system, including 12 instances of hand-raising, 18 of head-lowering, 7 of face-turning, and 10 of standing. Only 3 students had misclassified behaviors due to occlusion or similar factors.

4.2 Recognition Response Time Analysis

For response time, the system's average detection time per image is around 0.1s, and the video detection frame rate exceeds 25fps, meeting real-time monitoring requirements. In actual classroom teaching, when students are in class, after the teacher uploads a photo or turns on the camera, results can be obtained within a short time,

allowing real-time awareness of student engagement.

4.3 System Generalization Capability Analysis

To evaluate the system's generalization across various scenarios, experiments were conducted in dimly lit ordinary classrooms, crowded classrooms, and classrooms with background interference. The results show that the system maintains high recognition accuracy in all settings, demonstrating strong generalization. For instance, under low-light conditions, the system's accuracy in recognizing hand-raising behavior exceeds 90%, offering a clear advantage over traditional methods [6].

The experimental results above show that the system's use of deep learning for recognizing students' classroom behaviors meets the requirements for in-class behavior recognition in terms of accuracy, response time, and generalization performance. This system can provide technical support for classroom management.

5. RESULTS AND DISCUSSION

5.1 Experimental Results

This experiment implements a deep-learning-based behavior analysis of students' classroom learning activities. Tests and experiments demonstrate its strong feasibility and practicality, enabling accurate recognition of common classroom behaviors such as raising a hand, lowering the head, turning the head, and standing up, with an overall accuracy of 94.9%. In real classrooms, the recognition accuracy for student behaviors exceeds 93%, indicating significant practical application value.

5.2 Experimental Discussion

The development of this system carries significant value and innovation for educational and instructional use. By employing a deep-learning-based student classroom behavior recognition system, it enables automated and intelligent analysis of student behavior, eliminating the traditional reliance on teachers' subjective observations of classroom learning states. It provides teachers with objective, accurate data on students' classroom learning conditions, helping them gain deeper insight into student performance and promptly refine and adjust teaching methods to improve instructional quality. Moreover, the system supports real-time statistical analysis of student classroom behavior and presents the results in data tables. From these tables, teachers can deliver targeted instructional guidance based on observed behaviors, achieving personalized teaching. For example, teachers can use the system's statistics to learn how many students spoke in class or how many appeared distracted, and then address these issues accordingly [7].

6. CONCLUSION

In summary, the paper investigates the development of a classroom behavior recognition system based on deep learning. The above analysis shows that such a system offers a new pathway for education and teaching. The paper argues that the scope of behavior recognition should be expanded, algorithmic performance optimized, privacy protection strengthened, and system functions continuously improved to advance the in-depth application of deep-learning technology in education, laying a solid foundation for personalized learning and higher teaching quality.

REFERENCES

[1] Gao W and Gorinevsky D 2020 Probabilistic modeling for optimization of resource mix with variable generation and storage IEEE Trans. Power Syst. 35 4036–45
[2] Guo, Y. (2025). The Optimal Trajectory Control Using Deterministic Artifi cial Intelligence for Robotic Manipulator. Industrial Technology Research, 2(3).
[3] Guo, Y. (2025, May). IMUs Based Real-Time Data Completion for Motion Recognition With LSTM. In Forum on Research and Innovation Management (Vol. 3, No. 6).
[4] Guo, Y., & Tao, D. (2025). Modeling and Simulation Analysis of Robot Environmental Interaction. Artificial Intelligence Technology Research, 2(8).
[5] Wei, Xiangang, et al. "AI driven intelligent health management systems in telemedicine: An applied research study." Journal of Computer Science and Frontier Technologies 1.2 (2025): 78-86.

[6] Zhang, Yujun, et al. "MamNet: A Novel Hybrid Model for Time-Series Forecasting and Frequency Pattern Analysis in Network Traffic." arXiv preprint arXiv:2507.00304 (2025).
[7] Peng, Qucheng, Ce Zheng, and Chen Chen. "A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[8] Peng, Qucheng, et al. "RAIN: regularization on input and network for black-box domain adaptation." Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. 2023.
[9] Chen, J., Zhang, X., Wu, Y., Ghosh, S., Natarajan, P., Chang, S. F., & Allebach, J. (2022). One-stage object referring with gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5021-5030).
[10] Zhang, Jingbo, et al. "AI-Driven Sales Forecasting in the Gaming Industry: Machine Learning-Based Advertising Market Trend Analysis and Key Feature Mining." (2025).
[11] Zhang, Yuhan. "CrossPlatformStack: Enabling High Availability and Safe Deployment for Products Across Meta Services." (2025).
[12] Hu, Xiao. "AdPercept: Visual Saliency and Attention Modeling in Ad 3D Design." (2025).
[13] Li, X., Lin, Y., & Zhang, Y. (2025). A Privacy-Preserving Framework for Advertising Personalization Incorporating Federated Learning and Differential Privacy. arXiv preprint arXiv:2507.12098.
[14] Tu, Tongwei. "AutoNetTest: A Platform-Aware Framework for Intelligent 5G Network Test Automation and Issue Diagnosis." (2025).
[15] Zhu, Bingxin. "ReliBridge: Scalable LLM-Based Backbone for Small Business Platform Stability." (2025).
[16] Hu, Xiao. "Learning to Animate: Few-Shot Neural Editors for 3D SMEs." (2025).
[17] Tan, C., Gao, F., Song, C., Xu, M., Li, Y., & Ma, H. (2024). Proposed Damage Detection and Isolation from Limited Experimental Data Based on a Deep Transfer Learning and an Ensemble Learning Classifier.
[18] Xie, Minhui, and Shujian Chen. "Maestro: Multi-Agent Enhanced System for Task Recognition and Optimization in Manufacturing Lines." Authorea Preprints (2025).

Author Profile

Linying Yan (1990–), female, M.S., lecturer; research interests: deep learning, image recognition.