WHIOCE PUBLISHING PTE. LTD.
PROVIDING
FIRST-CLASS SCIENTIFIC INFORMATION
FOR TOP SCHOLARS

2025 Volume 3, Issue 6
Contemporary Education Frontiers
ISSN (Online): 3029-1860 ISSN (Print): 3029-1879

Analysis and Research on Students' Classroom Behavior Data Based on Object Detection

Baiyu Chen[1], **Deng Bian**[2], **Mingwei Tang**[2]*, **Mingfeng Zhao**[3]

[1]Library, Xihua University, Chengdu 610000, Sichuan, China

[2]School of Computer and Software Engineering, Xihua University, Chengdu 610039, Sichuan, China

[3]China Mobile Group Design Institute Co., Ltd. Sichuan Branch, Chengdu 610045, Sichuan, China

**Author to whom correspondence should be addressed.*

Abstract: Analyzing and studying students' classroom behavior is crucial for enhancing both students' abilities and teachers' instructional methods. This topic has been a significant focus within the educational community. Recently, machine vision and object detection technologies have been extensively applied across various domains, yielding notable outcomes. Consequently, this paper introduces a method for modeling and analyzing classroom behavior data using an object detection neural network. Experimental results indicate that this approach can effectively facilitate the development of students' abilities and the improvement of teaching practices.

Keywords: Component; Object Detection; Students' classroom behavior; Deep learning model

Online publication: July 26, 2025

1. Introduction

Students' classroom behavior significantly influences their personal growth and development, as well as teachers' effectiveness, school management, and the overall educational quality of society. Firstly, students' classroom behavior is closely linked to their academic performance and personal growth. Positive behaviors, such as attentive listening, active participation in discussions, and timely homework completion, enhance their understanding and mastery of knowledge. Secondly, students' behavior directly impacts teachers' teaching effectiveness and job satisfaction. Good classroom discipline and a positive, interactive learning environment enable teachers to impart knowledge and skills more efficiently, facilitating the achievement of teaching objectives. Moreover, students' classroom behavior reflects the educational quality and management standards of the school. Consistently good classroom behavior enhances the school's reputation and attractiveness, drawing in high-quality teachers and students. Lastly, from a broader societal perspective, good classroom behavior not only improves individual student quality but also fosters future citizens with a sense of responsibility, cooperation, and innovation. Improved educational quality has a long-term positive effect on societal stability and progress, contributing to a harmonious and orderly social environment.

Various schemes have been employed to analyze or monitor students' classroom behavior, such as the approach by

Buniyamin et al.[1], who classified and predicted students' scores using data from a campus information system. With the implementation of the smart campus project and the rapid advancement of computer technology, it has become feasible to acquire, process, and analyze classroom image data. Consequently, increasing numbers of researchers are focusing on analyzing students' classroom behaviors through this data to better assist teaching and management. The introduction of these new technologies has significantly changed the management and evaluation of classroom teaching. China's State Council has proposed utilizing intelligent technologies to accelerate the reform of teaching methods and talent cultivation models, develop intelligent educational assistants, and promote the application of AI in teaching, research, management, and other fields[2].

In recent years, artificial intelligence and deep learning have achieved significant breakthroughs and garnered substantial public attention. For example, the deep convolutional network proposed by LeCun, Bengio, and Hinton[3] is pivotal in the field of image recognition. Computer vision remains a crucial research direction in artificial intelligence because vision is the primary sensory source for information acquisition in humans. Treicher et al.[4] confirmed through experiments that visual information accounts for 83% of all information acquired by humans. Consequently, computer vision has become a major research focus for both academic institutions and enterprises, boasting the longest research history and the most extensive technological accumulation in the field of artificial intelligence.

Object detection techniques have diverse applications in computer vision, focusing on algorithms to identify and localize specific targets in images or videos. In classroom settings, these targets typically include students, teachers, and other relevant objects. Recently, the advent of deep learning algorithms, particularly Convolutional Neural Networks (CNN), has significantly enhanced the accuracy and efficiency of object detection. These technological advancements present new opportunities for classroom behavior detection.

CNN can directly extract and learn features from a dataset, automatically identifying hierarchical features from images. Through multi-level convolutional operations, CNN can capture low-level edge features and high-level semantic features, allowing the model to comprehensively understand and interpret image content. Specifically, the advantage of CNN lies in their ability to automatically identify key features in data without relying on manually designed feature extraction methods.

Therefore, this paper proposes a student classroom behavior detection model based on a convolutional neural network (CNN). The model employs a customized feature aggregation module and a feature propagation mechanism, enabling each scale feature to incorporate detailed contextual information, which facilitates more accurate detection and classification of targets. Specifically, the feature aggregation module accepts inputs from three scales of features and utilizes parallel deep convolutions to capture rich cross-scale information. Additionally, the propagation mechanism distributes features with rich contextual information across each detection scale. This approach effectively integrates feature information from different scales, enhances feature expression capability, and improves the model's detection performance. Consequently, the model can more accurately detect student behavior in the classroom, providing more reliable support and references for educators.

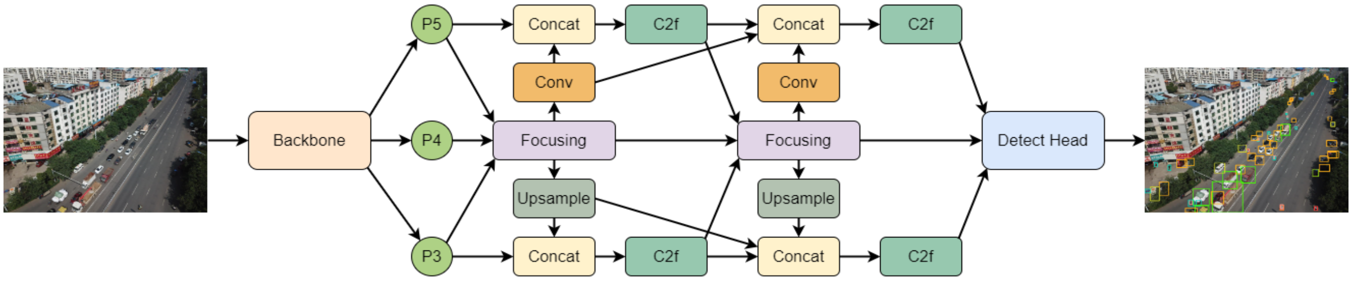


Figure 1. Model Framework

Figure 1. Model Framework

2. Related work and theory

2.1. Related Work

Object detection serves as the foundation for more complex computer vision tasks. Deep learning-based object detection methods can generally be categorized into one-stage and two-stage methods. One-stage methods[5-7] generate predictions from the image in a single seamless step, while two-stage methods [8-10] complete the task in two stages, progressing from coarse to fine. One-stage methods utilize anchor boxes on the feature map to predict object categories and coordinates. In contrast, two-stage methods initially generate high-quality region proposals through architectures like the Region Proposal Network (RPN) [8], and then the detection head uses these region features for subsequent object classification and localization. Due to the absence of proposals, one-stage methods are highly computationally efficient but lag in accuracy compared to two-stage methods. Anchor-free methods eliminate anchor boxes[11-14]. Instead of using predefined anchor boxes, these methods regress the object's center and dimensions directly on feature maps at different scales. This approach can help avoid issues such as missed detections or duplicate detections thatarise from improperly setting anchor boxes. Additionally, Carion et al.[15] first proposed DETR, an end-to-end detector based on Transformers, which eliminates the anchor mechanism and the NMS component and employs a two-end matching mechanism to directly predict the one-to-one object set. Subsequently, many variants of DETR[16-19] have been proposed to further optimize its convergence speed and computational cost.

Cross-scale Feature Fusion involves integrating features from different scales within an object detection model to enhance its ability to detect objects of varying sizes. This method is commonly employed in modern object detection networks[7,8,20]. In a typical convolutional neural network (CNN), features are extracted at different layers, each capturing distinct receptive fields and levels of semantic information. Shallow layers, closer to the input, preserve more spatial detail, while deeper layers, nearer the output, capture richer semantic content. By combining features across these layers, both spatial detail and semantic information are retained, leading to improved detection accuracy[21-25]. Cross-scale feature fusion is particularly effective in handling images with significant scale variations, complex backgrounds, and diverse object types.

2.2. Theory

For the input image data, the model's backbone network progressively extracts feature maps at different scales, generating a set of multi-layer feature maps denoted as [P3, P4, P5]. Each layer has distinct resolution and receptive field characteristics. We define a focusing module to integrate these feature maps across different layers. First, we apply upsampling (or downsampling) and convolution to adjust the size and number of channels in the feature maps, facilitating subsequent feature concatenation. Adjusting the number of channels ensures that essential information is retained during multi-scale feature fusion. Next, we concatenate the feature maps from different layers. This multi-scale feature fusion allows the model to comprehensively capture information at various scales. Following this, several deep convolutional layers process the concatenated features, and the outputs from different convolutional kernel sizes are summed to produce a feature map that incorporates multi-scale information. Finally, the feature map is passed through a pointwise convolutional layer, which further fuses and compresses the channels to maintain compact and efficient features. Additionally, we employ residual connections in the module: the processed feature maps from the deep convolutional layers are added to the original concatenated feature maps, forming a residual structure. This design mitigates the vanishing gradient problem in deep networks and accelerates convergence. The focusing module equation can be expressed as:

$$P_{in}=ADown_{3\times3}(P_3)+Conv_{1\times1}(P_4)+Conv_{1\times1}\big(Upsample(P_5)\big),$$

$$P_{mid}^{(m)}=DWConv_{k^{(m)}\times k^{(m)}}(P_{in}),\ m=1,\ldots,4,$$

$$P_{out}=Conv_{1\times1}\left(\sum_{m=1}^{4}P_{mid}^{(m)}+P_{in}\right)+P_{in}.$$

The feature map generated after the focusing module contains detailed contextual information. To propagate this information across different detection scales, we design a feature diffusion network. The feature maps are resized using upsampling or downsampling operations and then concatenated with the original hierarchical feature maps to transfer the contextual information to the corresponding levels. These concatenated maps are subsequently processed through the C2f module, based on a CSP structure [26], to enhance their expressive power. Following this, the focusing module integrates features from three different scales. The resulting feature maps are again resized and concatenated with corresponding hierarchical feature maps, and the refined maps are passed through the C2f module to perform the final detection tasks.

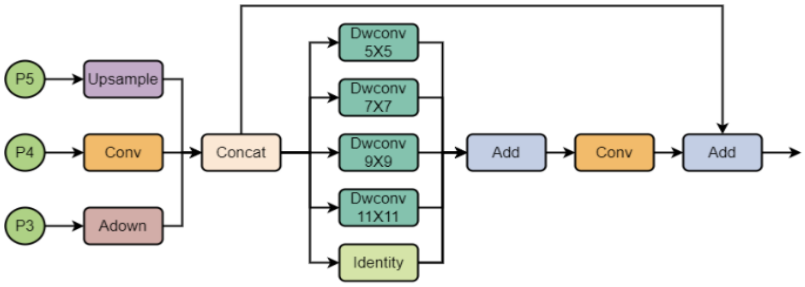


Figure 2. Focusing Model

The loss function quantifies the difference between the model's predicted results and the actual outcomes, serving as the objective for the optimization algorithm. In this paper, the loss function consists of two components: classification loss and regression loss. The classification loss is calculated using a modified version of the binary cross-entropy (BCE) loss, which is expressed by the following formula:

$$L_{BCE}=\frac{1}{N}\sum_{i=1}-[t_i\cdot\log(p_i)+(1-t_i)\cdot\log\boxed{f0}(1-p_i)]$$

where N represents the object categories, a single encoding of the true label representing the true probability of i categories, and is the probability of i categories predicted by the model.

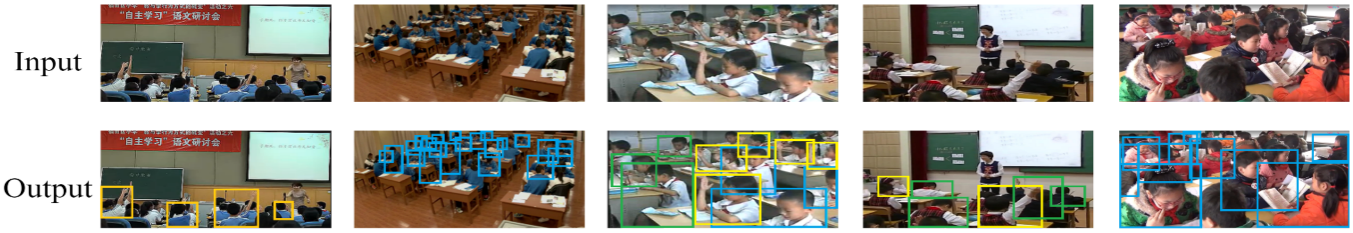
The regression loss function includes Complete Intersection over Union (CIoU) loss and Distribution Focal Loss (DFL). First, DFL is used to compute the loss between the bounding box distribution probability and the label distribution probability, optimizing each edge of the predicted box. Next, the bounding box distribution probability is reduced to the predicted box, and the CIoU loss calculates the difference between the predicted box and the ground truth box, optimizing the prediction as a whole:

$$L_{DFL}(Q_i,Q_{i+1})=-((t_{i+1}-t)\log(Q_i)+(t-t_i)\log\boxed{f0}(Q_{i+1}))$$

Here, $Q_i=\frac{t_{i+1}-t}{t_{i+1}-t_i}$ and $Q_{i+1}=\frac{t-t_i}{t_{i+1}-t_i}$ represent the prediction scores for a fixed distance from the center of the target to the edge of the predicted box, ensuring that the estimated regression target approaches the corresponding label t as closely as possible. IoU (Intersection over Union) [27] denotes the ratio of the intersection area to the union area of two boxes. e represents the Euclidean distance between the two boxes, while b and b^{gt} represent the centroids of the predicted and ground truth boxes, respectively. d denotes the diagonal distance between the closest regions of the two boxes. r is the weighting coefficient, and v measures the consistency of the relative proportions of the two boxes.

The loss function of the algorithm is obtained by weighted summation of the three loss functions:

$$L_{CIoU}=1\text{-}IoU+\frac{e^2(b,b^{gt})}{d^2}+rv$$



$$L_{Total}=w_1L_{BCE}+w_2L_{DFL}+w_3L_{CIoU}$$

Figure 3. Detection effect of the model

3. Experimental settings

The input image is configured as 640x640, the batch size is set to 8 and there are no pre-training weights. To address the overfitting problem, we used an early stopping strategy with a patience value of 20. A stochastic gradient descent (SGD) optimizer was used by default with a learning rate of 0.01. The learning rate was 0.01. We followed the training of 100 epochs as recommended in the official guidelines[28].

4. Experimental results and data analysis

In this section, to demonstrate the superiority of our method, we compared it with the latest state-of-the-art techniques, as shown in **Table 1**. Compared to multiple versions of the YOLO algorithm (a widely-used object detection framework), our algorithm outperforms others in recognizing three specific types of student behaviors (e.g., read, write, and raise hand). Additionally, our method achieved a mean Average Precision (mAP) of 73.0%.

Figure 3 provides a visual representation of the algorithm's ability to detect and classify student behaviors in the classroom. The bounding boxes in **Figure 3** indicate the detected objects, with their quantity reflecting the algorithm's detection capacity and their alignment with object boundaries demonstrating its accuracy. Furthermore, the different colors of the boxes correspond to different student behaviors, making it easier to visually distinguish between behavior categories[29].

It can be observed that our algorithm detects a greater number of target objects, and it correctly identifies the categories of student behaviors in the classroom. Therefore, both the comparative experiments and the visual results demonstrate the high accuracy and reliability of our method in recognizing student behaviors in the classroom[30].

Table 1. Comparison with advanced methods

Method	**Class**			**mAP@0.5**
	Rise Hand	**Write**	**Read**	
Yolov5s[20]	72.6%	71.5%	57.8%	67.3%
Yolov8s[20]	71.7%	74.5%	58.9%	68.7%
Yolov9s [28]	81.4%	74.9%	59.7%	72.0%
Yolov10s[29]	78.5%	72.9%	59.4%	70.3%
Yolov11s[30]	80.2%	74.3%	58.2%	70.9%
Ours	82.4%	75.5%	61.2%	73.0%

5. Conclusion

The paper proposes an analysis of and research into students' classroom behavior data based on an improved object detection method. It has been demonstrated that the proposed model performs exceptionally well in experiments. Our proposed method is significant for assessing and enhancing the quality of students' classroom learning as well as the quality of teachers' teaching. It enhances students' focus while simultaneously supporting teachers in refining their instructional strategies. Future work will explore more effective motion detection modules.

Funding

This work is supported by the Scientific Research Funds project of Science and Technology Department of Sichuan Province (Project No.: 2019YFG0508,2019GFW131, 2023), the National Natural Science Foundation of China (Project No.: No. 61902324), Funds Project of Chengdu Science and Technology Bureau (Project No.: 2017-RK00-00026-ZF,) and Science, Technology Planning Project of Guizhou Province (Project No.: QianKeHeJiChu-ZK[2021]YiBan319), the Xihua University Education and teaching reform project (Project No.: xjjg2021049, xjjg2021115) and the Sichuan Province's Higher Education Talent Training and Teaching Reform Project for 2021-2023 (Project No.: JG2021-929).

Disclosure statement

The author declares no conflict of interest.

References

[1] Buniyamin N, Mat U, Arshad P M, 2015, Educational data mining for prediction and classification of engineering students achievement. //2015 IEEE 7th International Conference on Engineering Education (ICEED). IEEE: 49-53.

[2] Lu G, Xie K, Liu Q, 2021, Automated Annotation of Classroom Behaviours with AI Engine. Open Education Research, 27(06): 97-107.

[3] Le C Y, Bengio Y, Hinton G, 2015, Deep learning. nature, 521(7553): 436-444.

[4] Treichler D G, 1967, Are you missing the boat in training aids. Film and AV Communication, 1: 14-16.

[5] Lin T Y, Goyal P, Girshick R, et al., 2017, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2980–2988.

[6] Redmon J, Divvala S, Girshick R, et al., 2016, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 779–788.

[7] Liu W, Anguelov D, Erhan D, et al., 2016, Ssd: Single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, Proceedings, Part I 14, Springer, 21–37.

[8] Ren S, He K, Girshick R, et al., 2017, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE transactions on pattern analysis and machine intelligence, 39(6): 1137-1149.

[9] Girshick R, 2015, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 1440–1448.

[10] He K, Zhang X, Ren S, et al., 2015, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE transactions on pattern analysis and machine intelligence, 37(9): 1904-1916.

[11] Tian Z, Shen C, Chen H, et al., 2022, Fcos: A simple and strong anchor-free object detector, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(4): 1922-1933.

[12] Zhou X, Wang D, Kr¨ahenb¨uhl P, 2019, Objects as points, arXiv preprint arXiv: 1904.07850.

[13] Duan K, Bai S, Xie L, et al., 2019, Centernet: Keypoint triplets for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision: 6569-6578.

[14] Law H, Deng J, 2018, Cornernet: Detecting objects as paired keypoints, in: Proceedings of the European conference on computer vision (ECCV): 734–750.

[15] Carion N, Massa F, Synnaeve G, et al., 2020, End-to-end object detection with transformers, in: European conference on computer vision, Springer: 213–229.

[16] Liu S, Li F, Zhang H, et al., 2022, DAB-DETR: Dynamic anchor boxes are better queries for DETR, in: International Conference on Learning Representations. https: //openreview.net/forum?id=oMI9PjOb9Jl.

[17] Li F, Zhang H, Liu S, et al., 2022, Dn-detr: Accelerate detr training by introducing query denoising, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition: 13619–13627.

[18] Roh B, Shin J, Shin W, et al., 2021, Sparse detr: Efficient end-to-end object detection with learnable sparsity, arXiv preprint arXiv: 2111.14330.

[19] Chen Q, Chen X, Zeng G, et al., 2022, Group detr: Fast training convergence with decoupled one-to-many label assignment, arXiv preprint arXiv: 2207.2207.13085.

[20] Jocher G, Qiu J, Chaurasia A, 2023, Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics.

[21] Woo S, Hwang S, Kweon I S, 2018, Stairnet: Top-down semantic aggregation for accurate one-shot detection. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE: 1093-1102.

[22] Liu Z, Gao G, Sun L, 2020, Ipg-net: Image pyramid guidance network for small object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops: 1026-1027.

[23] Gong Y, Yu X, Ding Y, 2021, Effective fusion factor in fpn for tiny object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision: 1160-1168.

[24] Hong M, Li S, Yang Y, 2022, Sspnet: Scale selection pyramid network for tiny person detection from uav images. IEEE geoscience and remote sensing letters, 19: 1-5.

[25] Li S, Huang D, Wang Y, 2019, Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:1911.09516.

[26] Wang C Y, Liao H Y M, Wu Y H, 2020, CSPNet: A new backbone that can enhance learning capability of CNN// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 390-391.

[27] Yu J, Jiang Y, Wang Z, 2016, Unitbox: An advanced object detection network//Proceedings of the 24th ACM international conference on Multimedia. 516-520.

[28] Wang C Y, Yeh I H, Mark L H Y, 2025, Yolov9: Learning what you want to learn using programmable gradient information//European Conference on Computer Vision. Springer, Cham, 1-21.

[29] Wang A, Chen H, Liu L, 2024, Yolov10: Real-time end-to-end object detection. arxiv preprint arxiv: 2405.14458.

[30] Khanam R, Hussain M, 2024, Yolov11: An overview of the key architectural enhancements. arxiv preprint arxiv:2410.17725.

Publisher’s note

Whioce Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.