A Computer Vision-Based Approach for Assessing Student Behavioral Engagement in the Classroom

Yaping XU[a,b], Haogang BAO[c], Yaqian ZHENG[d], Siyu WANG[a], Sisi WU[a], & Yanyan LI [a,*]

[a]*Faculty of Education, Beijing Normal University, Beijing, China*
[b]*School of Educational Technology, Northwest Normal University, Lanzhou, China*
[c]*The China National Academy of Educational Sciences, Beijing, China*
[d] *College of Education, Hebei University, Baoding, China*
*liyy@bnu.edu.cn

Abstract: With increasing emphasis on enhancing classroom engagement and learning outcomes, efficient and accurate assessment of student behavioral engagement has become essential. Traditional methods such as classroom observations and self-reports are often subjective, time-consuming, and lack scalability. To address these limitations, this study proposes a computer vision-based approach for automatically assessing student behavioral engagement in classroom environment. Specifically, an improved VGG16-Attn model is proposed by integrating attention mechanisms to enhance feature extraction and boost behavior recognition performance. Our results indicate that the proposed method not only accurately detects student behavior types but also, by further aggregating and calculating these behaviors, objectively and effectively captures the dynamic evolution of behavioral engagement. This method offers a promising solution for educators to gain actionable insights into student engagement, ultimately contributing to more personalized and effective teaching strategies.

Keywords: Behavioral engagement, computer vision, classroom environment

1. Introduction

Student engagement can be defined as the degree to which students participate in learning activities, encompassing their attention, involvement, and effort (Fredricks et al., 2004). Research has shown a significant positive correlation between student engagement and academic achievement (Gunuc, 2014). Student engagement is conceptualized as a multidimensional construct consisting of cognitive, emotional, and behavioral components (Fredricks et al., 2004; Cooper, 2014), where cognitive and emotional engagement are relatively implicit and difficult to observe directly. In contrast, behavioral engagement is more easily assessed as it is directly reflected in students' observable actions, such as classroom interactions, answering questions, and completing assignments. Therefore, behavioral engagement is often the most accessible dimension to evaluate and is the most commonly used indicator in many studies. Traditionally, engagement has been assessed through self-reports, teacher observations, and surveys, but these methods often suffer from subjectivity, lack of real-time feedback, and difficulties in scaling for large classrooms.

Recent advancements in computer vision and deep learning have provided new avenues for automating the assessment of student behavioral engagement in classroom environments. These technologies offer the potential to capture real-time behavioral data with greater accuracy and consistency compared to traditional methods. A growing body of research has demonstrated the potential of using video-based analysis for tracking student behaviors (Ngoc Anh et al., 2019; Guo, 2022; Trabelsi et al., 2023). Despite these advancements, video-based behavioral analysis still faces several challenges in practical applications. First, the performance of behavior recognition models is limited by the scale and diversity of available datasets. Many datasets fail to adequately cover various classroom environments and behavior types. Second, existing research often lacks long-term tracking of student behaviors, which hinders a comprehensive analysis of the evolution of students' behavioral engagement. Additionally, most studies primarily focus on the technical aspects of behavior recognition. While these studies have achieved relatively accurate recognition of student behaviors, they have not effectively integrated this behavioral data into a comprehensive assessment of behavioral engagement.

To address the aforementioned challenges, this study constructed a large-scale, high-quality dataset specifically designed to support the task of student behavior recognition in classroom contexts. On the basis of this dataset, a computer vision–driven framework was proposed for the quantitative assessment of students' behavioral engagement during live classroom sessions. This framework integrates advanced visual analysis techniques to capture and interpret students' actions, enabling a more objective and scalable measurement of engagement levels. The findings of this study contribute to providing teachers with accurate classroom feedback and offer data support for optimizing teaching strategies and implementing personalized education.

2. Literature Review

2.1 Behavioral Engagement and Its Typical Measurement

The definition of classroom behavioral engagement can be categorized into two main perspectives: one focuses on student compliance with rules, specifically reflecting behaviors such as attendance rates. For instance, Finn (1989) suggests that learning engagement helps identify the process of students gradually becoming disengaged and alienated from school, thus enabling timely interventions to support students in completing their academic work. The other perspective examines the extent of students' deep involvement in learning activities, including effort, persistence, and concentration (Fredricks et al., 2004). This is typically reflected in behaviors such as completing assignments, attending classes, answering questions, and participating in discussions. This study, building upon the foundation of prior research, conceptualizes behavioral engagement as the proactive and purposeful interaction of students with various learning resources, instructors, and peers within the classroom context. Such engagement encompasses actions that reflect students' participation, persistence, and effort in academic tasks. By systematically observing students' behavioral states throughout the learning process, the degree of their engagement in classroom activities can be assessed.

Traditional methods for measuring classroom behavioral engagement primarily include self-reports and classroom observations. Early research primarily involved learners self-reporting their levels of engagement. This method is highly practical and provides large-scale data at a relatively low cost. However, self-reports are prone to subjective bias, and students may not always provide truthful responses. Additionally, this approach fails to offer process-oriented analysis, making it difficult to track the dynamic changes in student engagement over time. Classroom observation, on the other hand, uses pre-established observation scales to capture students' behavioral characteristics during the learning process, thereby assessing their behavioral engagement (Volpe et al., 2005). The primary limitation of this method lies in the substantial investment of time and effort required from researchers, rendering it impractical for large-scale investigations and unsuitable as a routine instrument for process-oriented assessment. However, the advent of computer vision technology has introduced new opportunities to enhance and transform traditional classroom observation approaches. Leveraging cameras, sensors, and sophisticated algorithms, computer vision systems are capable of capturing and processing visual data in real time, thereby offering researchers granular and dynamic insights into student behaviors and interactions. In particular, recent advancements in deep learning have significantly expanded the analytical capabilities of these systems, enabling more accurate, scalable, and automated behavioral analysis. As a result, research integrating computer vision with educational observation has attracted growing scholarly interest in recent years, paving the way for more efficient and data-rich methods of studying classroom engagement.

2.2 Classroom Behavioral Engagement Recognition via Computer Vision

Computer vision technology aims to detect, recognize, track, and understand objects through the processing and analysis of image or video data. Researchers can utilize this technology to perform fine-grained recognition of students' body movements in classroom videos, enabling the automated assessment of behavioral engagement. Early studies in this field primarily focused on using traditional machine learning methods to recognize classroom engagement (Karimah & Hasegawa, 2022). For instance, Zaletelj et al. (2017) trained a classifier to automatically assess students' attention levels in the classroom by extracting features from facial expressions and body posture, and combining them with various traditional machine learning algorithms. The results indicated that the three-level attention classifier achieved a moderate accuracy of 75.3%. The advantage of traditional machine learning methods lies in their relatively simple model structures, lower computational requirements, and suitability for smaller datasets. However, these methods are highly dependent on manual feature extraction and have limited effectiveness in handling complex and dynamic student behaviors.

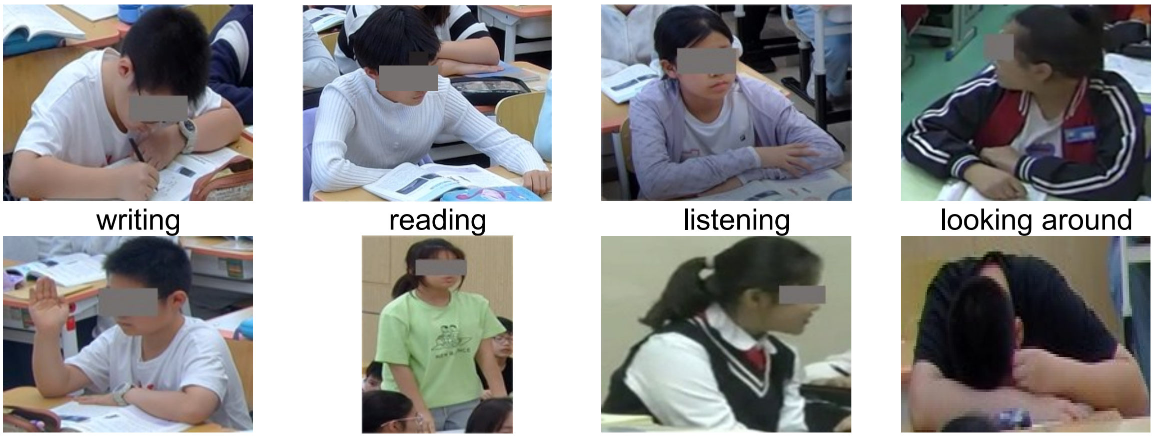
With the advancement of artificial intelligence, deep learning methods have demonstrated their superiority in image recognition tasks. The end-to-end learning process not only simplifies the workflow of traditional machine learning tasks but also enhances the capability of deep neural networks through a hierarchical feature extraction process. Consequently, in recent years, an increasing number of studies on student behavior recognition have adopted deep learning methods. For example, Li et al. (2023) used classroom videos from a smart classroom to generate an image dataset containing seven typical learning behaviors. Based on this dataset, they developed a behavior recognition network model, which includes a backbone network (SlowFast R-101) for feature extraction, a target detector (Faster R-CNN), and an action classifier composed of fully connected layers. Ikram et al. (2023) employed a VGGNet16 model based on transfer learning to calculate students' learning engagement levels in real classroom settings by extracting external behavioral features. Additionally, Xiong et al. (2024) proposed a CNN-Transformer model that simultaneously captures both coarse- and fine-grained information, significantly improving the accuracy of learning engagement recognition in classroom contexts.

In summary, research on classroom behavioral engagement recognition based on computer vision has evolved from traditional methods based on handcrafted features to deep learning approaches. Although deep learning methods are widely favored for their ability to provide more accurate recognition of student behavior, they often suffer from lower interpretability and typically require large amounts of labeled data for training. Existing datasets in current research are relatively limited in size, which can lead to overfitting issues. Furthermore, due to privacy concerns and other factors, acquiring datasets from external institutions presents significant challenges. In response to these issues, this study has constructed a large-scale student behavior dataset by collecting real classroom videos and has conducted an in-depth analysis of engagement levels based on behavior recognition results.

3. Methodology

3.1 Student Behavior Dataset

Due to the absence of publicly available datasets tailored to real-world classroom settings, we constructed a large-scale student behavior dataset to support behavior recognition model development and evaluation. The dataset includes 152,823 annotated images extracted from actual classroom videos, each labeled with one of nine representative behaviors: listening, reading, writing, raising hand, standing, leaning the body, looking around, lying on desk, and other, Figure 1 shows example images. These categories were selected for their pedagogical relevance and importance in evaluating student engagement. The dataset captures diverse classroom environments, camera angles, postures, and occlusion conditions, making it highly representative. This resource provides a solid foundation for advancing computer vision-based behavior analysis in education.



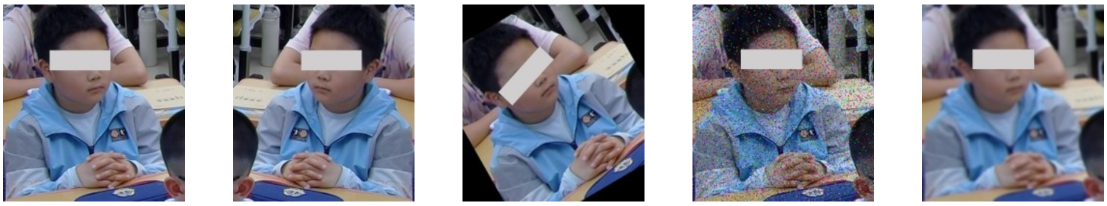
writing reading listening looking around

raising hand standing leaning the body lying on desk

Figure 1. Example images from the behavior dataset.

3.2 Data Augmentation

The annotated behavior dataset exhibits class imbalance, which may lead to biased predictions if directly used to train deep learning models. In such cases, the model is likely to favor the majority class, resulting in poor recognition performance for the minority classes. To expand and balance the image dataset, this study employed various data augmentation (DA) techniques, including rotation, flipping, noise addition, and blurring, as shown in Figure 2. After applying these augmentation methods, the final dataset used to train the behavior recognition model contained 312,459 images.



Original Image Flipping Rotation Noise Addition Blurring

Figure 2. Some examples of DA.

3.3 Improved VGGNet16 for Recognizing Student Behavior

Although VGGNet16 (Simonyan & Zisserman, 2019) is a classic convolutional neural network with strong performance in image classification, it lacks the ability to adaptively focus on important spatial and channel-wise features. This limitation reduces its effectiveness in distinguishing key behavioral cues from complex classroom backgrounds, potentially lowering recognition accuracy.

To overcome this, we propose an improved model, referred to as VGG16-Attn, by embedding a Convolutional Block Attention Module (CBAM) at the end of each convolutional block in the original VGGNet16 (see Figure 3). The CBAM module introduces a two-stage attention mechanism: the channel attention sub-module emphasizes meaningful feature channels by learning inter-channel relationships, while the spatial attention sub-module highlights relevant regions in the feature map by learning spatial dependencies. Structurally, the modified model retains the five convolutional blocks of VGGNet16, each followed by a CBAM module, and ends with three fully connected layers for classification. This enhancement enables the network to better capture discriminative features relevant to student behavior, improving both accuracy and robustness.

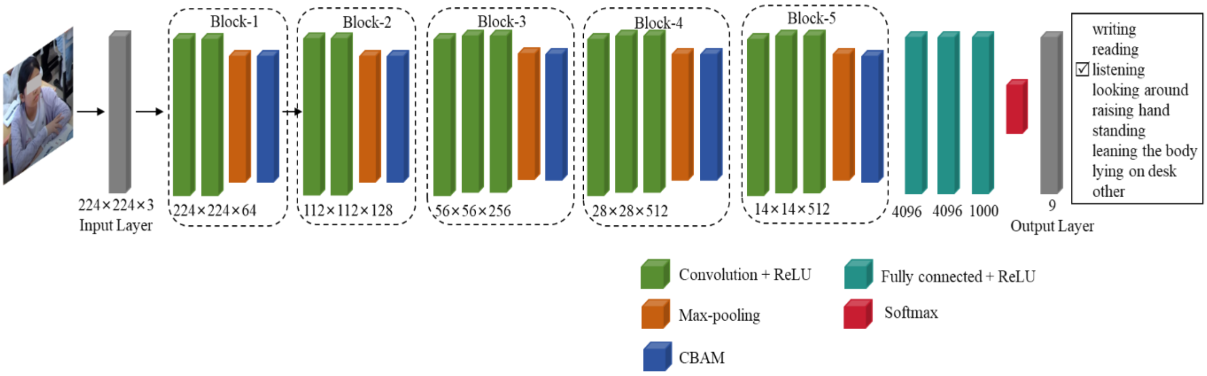


Figure 3. Network architecture of the VGG16-Attn model.

3.4 Behavioral Engagement Analysis

Chi and Wylie (2014) proposed the ICAP (Interactive, Constructive, Active, Passive) framework, which distinguishes students' engagement levels in the classroom based on their observable behaviors. This study builds on the ICAP framework to systematically categorize learning behaviors and classifies them into four types: interactive behavior, active behavior, passive behavior, and disengaged behavior. Specifically, interactive behavior emphasizes students' interaction with others (teacher or peers) in the classroom, and is demonstrated by actions such as standing up to answer questions or leaning the body to communicate. Active behavior reflects active participation and autonomous learning tendencies, such as writing or raising hands. Passive behavior refers to actions where students primarily receive information in the classroom, such as listening or reading. Disengaged behavior refers to a lack of meaningful learning participation, such as looking around or lying on the desk.

The levels of engagement for these four types of behavior are ranked from highest to lowest as follows: active behavior > interactive behavior > passive behavior > disengaged behavior. To quantitatively reflect this hierarchy, weights representing each behavior’s contribution to overall engagement are assigned using the Analytic Hierarchy Process (AHP) (Tavana et al., 2023), as shown in Table 1.

Table 1. *Weight Distribution Information for Different Behavior Types*

Behavior Type	Weight	Description	Explicit Indicators
Active behavior	0.4824	Actively participate in learning, but do not involve interaction.	writing, raising hand
Interactive behavior	0.3213	It Involves collaboration and knowledge-building activities.	standing, leaning the body
Passive behavior	0.1457	Limited to information reception, indicating a lower level of engagement.	listening, reading
Disengaged behavior	0.0506	Not engaged in classroom learning.	looking around, lying on desk, other

After obtaining the behavior recognition results for all students in the entire class using the trained model, the occurrence proportion of each behavior type within each time segment (5 seconds in this study) can be calculated. The behavior engagement level for each time segment can then be computed using the following formula:

$$Behavioral\ Engagement = \sum_{i=1}^{n} (W_i \times P_i)$$

In the formula, W_i represents the weight of the i-th behavior type, P_i denotes the proportion of occurrences of the i-th behavior type, and n refers to the total number of behavior types (n=4).

4. Result and Discussion

This section presents the baseline results obtained using the dataset. The preliminary evaluation offers insights into the dataset's usability. Additionally, the trained model is applied to recognize student behaviors in real classroom videos, facilitating an analysis of the evolution of behavioral engagement.

4.1 Experimental Setup and Evaluation Metrics

The proposed model was evaluated on the self-constructed student behavior dataset, split into training, validation, and test sets in a 7:1:2 ratio. Experiments were conducted using Python with Keras 2.2.0 and TensorFlow-GPU 2.6.2. The model was trained for 50 epochs with a batch size of 32 and a learning rate of 0.01. To address class imbalance, a weighted cross-entropy loss function was employed, and parameter optimization was carried out using the Adam optimizer.

The performance of the proposed model was rigorously evaluated using four widely recognized metrics—Accuracy, Precision, Recall, and F1-score—thereby ensuring a comprehensive assessment of its predictive capabilities. All metrics were computed on the test set using macro-averaging, which ensures equal weight across all behavior categories regardless of class size. This approach enables a fair evaluation of the model's performance on both majority and minority classes.

4.2 Student Behavior Recognition Results

To comprehensively evaluate the effectiveness of the proposed VGG16-Attn model in classroom student behavior recognition, a series of comparative experiments were conducted against several representative state-of-the-art deep learning architectures. Specifically, the comparison set included InceptionV3 (Szegedy et al., 2016), Xception (Chollet, 2017), DenseNet121 (Huang et al., 2017), MobileNetV3 (Howard et al., 2017), and the original VGGNet16 (Simonyan & Zisserman, 2014). These models are widely used in computer vision tasks and exhibit distinct architectural advantages: InceptionV3 employs parallel multi-scale convolution modules to enhance feature extraction; Xception uses depthwise separable convolutions to reduce computational cost; DenseNet121 promotes feature reuse across layers to improve gradient flow and parameter efficiency; MobileNetV3 combines neural architecture search with optimized activation functions for lightweight deployment; and VGGNet16, though structurally simple, remains effective in many visual tasks due to its stable convolutional stacking.

The detailed results are summarized in Table 2. The experimental results indicate that although the proposed VGG16-Attn model has a higher computational complexity than the baseline models, it delivers the best classification performance across all core metrics. Specifically, it achieves an Accuracy of 0.91, a Macro-Precision (Macro-P) of 0.91, a Macro-Recall (Macro-R) of 0.90, and a Macro-F1 score of 0.91. Compared to the original VGGNet16, the VGG16-Attn model demonstrates a 5% improvement in accuracy, confirming the effectiveness of the CBAM attention mechanism in enhancing feature representation. In contrast, InceptionV3, Xception, and DenseNet121 exhibit lower computational costs, with Xception performing slightly better in select metrics. However, none of these models achieve an accuracy above 0.85, suggesting limited capacity to capture fine-grained behavior patterns in complex classroom scenarios. Notably, MobileNetV3 achieves the lowest computational complexity (only 0.44G FLOPs), underscoring its efficiency. Nonetheless, its classification accuracy is over 4% lower than that of VGG16-Attn, indicating that lightweight models may face challenges in extracting sufficiently rich features for recognizing complex student behaviors.

Table 2. *Comparison Results Between the VGG16-Attn and State-of-the-art Models*

Models	Accuracy	Macro-P	Macro-R	Macro-F1	FLOPs(G)
InceptionV3	0.77	0.77	0.76	0.76	5.69
Xception	0.84	0.87	0.83	0.84	9.11
DenseNet121	0.70	0.71	0.67	0.68	5.67
MobileNetV3	0.87	0.86	0.87	0.87	0.44
VGGNet16	0.86	0.87	0.86	0.86	15.47
VGG16-Attn (This study)	0.91	0.91	0.90	0.91	30.95

Figure 4 displays the results of student behavior detection in different classroom scenarios using the trained model. As shown in the figure, the majority of students' behavior states are accurately recognized, indicating that the model performs well under ideal lighting conditions and with relatively clear visibility. However, for students seated at the back or in corner positions of the classroom, the model faces challenges in behavior recognition due to limited visibility and reduced resolution.

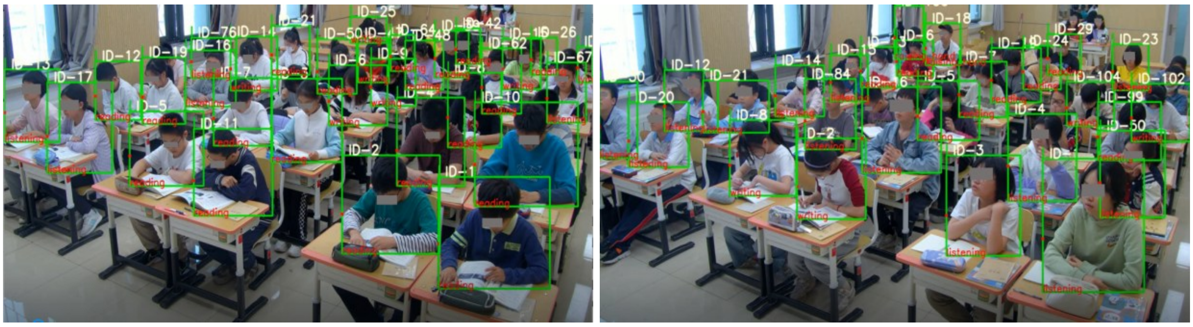


Figure 4. Examples of student behavior recognition.

4.3 Analysis of Classroom Behavioral Engagement

This study analyzed classroom behavioral engagement using video recordings from two class sessions. Prior to data collection, informed consent was obtained, and students were clearly informed about the purpose of the study. All data were anonymized to ensure student privacy and comply with ethical standards.

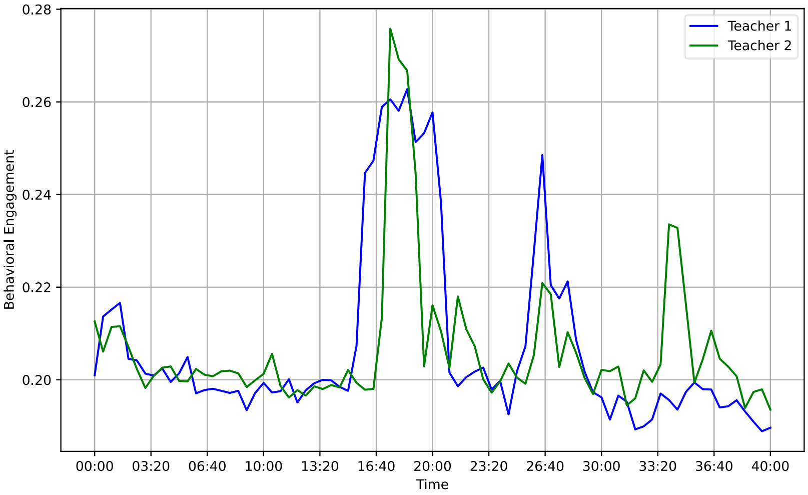


Figure 5. Temporal evolution of student behavioral engagement in two classrooms.

Aggregated behavior recognition results revealed distinct engagement patterns (Figure 5): both classes showed low engagement during the first 15 minutes, with noticeable increases around the 16th and 27th minutes. However, in the final 10 minutes, engagement diverged—remaining low in Teacher 1's class, while fluctuating in Teacher 2's class. Video analysis indicates that low engagement corresponded to passive lecture segments, whereas peaks aligned with interactive activities such as practice and questioning. Despite identical instructional content, the engagement trends reflect differences in teaching pace and style. Teacher 2's class included more frequent but shorter practice sessions, consistent with lesson evaluation notes stating that "the pace of the lesson was faster, and when students made mistakes, the teacher quickly moved on without targeted feedback". While this approach covered more content, it limited opportunities for in-depth learning. These findings are consistent with classroom observations and further validate the effectiveness of the proposed automatic behavioral engagement recognition method.

5. Conclusions and Future Work

This study integrates computer vision technology with an improved deep learning model to enable automated assessment of classroom behavioral engagement, aiming to address the limitations of traditional methods—namely, subjectivity, lack of real-time feedback, and poor scalability. Furthermore, the proposed approach was applied to real classroom video analysis, and the experimental results confirmed its effectiveness and reliability in practical settings. This provides a scientific foundation for teachers to dynamically monitor students' learning processes.

Although this study successfully achieves automatic recognition of classroom behavioral engagement using classroom video data, it has not yet focused on more implicit aspects of engagement, including emotion and cognition. Future research should integrate other sources of data, such as classroom dialogues, and further develop a more comprehensive, multimodal approach to automated learning engagement assessment. Additionally, future work should explore how to better integrate the proposed method into teaching practices to enhance educational efficiency and quality, thereby contributing to the development of educational informatization and intelligence.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No: 62277006, 62267008).

References

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition*, 1251-1258. https://doi.org/10.48550/arXiv.1610.02357

Cooper, K. S. (2014). Eliciting engagement in the high school classroom: A mixed-methods examination of teaching practices. *American Educational Research Journal*, 51(2), 363-402. https://www.jstor.org/stable/24546691

Finn, J. D. (1989). Withdrawing from school. Review of Educational Research, 59(2), 117-142. https://doi.org/10.2307/1170412

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59-109. https://doi.org/10.3102/00346543074001059

Gunuc, S. (2014). The relationships between student engagement and their academic achievement. *International Journal on New Trends in Education and their Implications*, 5(4), 216-231. https://doi.org/10.1007/s40299-013-0095-8

Guo, Q. (2022). System analysis of the learning behavior recognition system for students in a law classroom: Based on the improved SSD behavior recognition algorithm. *Scientific Programming*, 2022(1), 3525266. https://doi.org/10.1155/2022/3525266

Howard, A. G. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition*, 4700-4708. https://doi.org/10.48550/arXiv.1608.06993

Ikram, S., Ahmad, H., Mahmood, N., Faisal, C. N., Abbas, Q., Qureshi, I., & Hussain, A. (2023). Recognition of student engagement state in a classroom environment using deep and efficient transfer learning algorithm. *Applied Sciences*, 13(15), 8637. https://doi.org/10.3390/app13158637

Karimah, S. N., & Hasegawa, S. (2022). Automatic engagement estimation in smart education/learning settings: A systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments*, 9(1), 31. https://doi.org/10.1186/s40561-022-00212-y

Li, Y., Qi, X., Saudagar, A. K. J., Badshah, A. M., Muhammad, K., & Liu, S. (2023). Student behavior recognition for interaction detection in the classroom environment. *Image and Vision Computing*, 136, 104726. https://doi.org/10.1016/j.imavis.2023.104726

Ngoc Anh, B., Tung Son, N., Truong Lam, P., Phuong Chi, L., Huu Tuan, N., Cong Dat, N., ... & Van Dinh, T. (2019). A computer-vision based application for student behavior monitoring in classroom. *Applied Sciences*, 9(22), 4729. https://doi.org/10.3390/app9224729

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818-2826. https://doi.org/10.48550/arXiv.1512.00567

Tavana, M., Soltanifar, M., & Santos-Arteaga, F. J. (2023). Analytical hierarchy process: Revolution and evolution. *Annals of Operations Research*, 326(2), 879-907. https://doi.org/10.1007/s10479-021-04432-2

Trabelsi, Z., Alnajjar, F., Parambil, M. M. A., Gochoo, M., & Ali, L. (2023). Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. *Big Data and Cognitive Computing*, 7(1), 48. https://doi.org/10.3390/bdcc7010048

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review*, 34(4), 454-474. https://doi.org/10.1080/02796015.2005.12088009

Xiong, Y., Xinya, G., & Xu, J. (2024). CNN-Transformer: A deep learning method for automatically identifying learning engagement. *Education and Information Technologies*, 29(8), 9989-10008. https://doi.org/10.1007/s10639-023-12058-z

Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing*, 2017, 1-12. https://doi.org/10.1186/s13640-017-0228-8