**What is the main objective or research question of this study?**

The main objective of the study is to develop a bimodal learning engagement recognition method that automatically monitors and categorizes students' engagement levels in offline classroom settings using non-invasive video analysis. The research aims to recognize engagement through both emotional and behavioral cues, such as facial expressions, hand gestures, and body posture, to provide a comprehensive understanding of student engagement during lessons.

**What gap in the existing literature does this study aim to address?**

This study aims to address the gap in existing literature regarding the lack of comprehensive datasets and methods for recognizing student engagement in real classroom environments. Most current engagement recognition methods rely on single-modal approaches or limited datasets, which do not capture the complexity of student behavior in actual classroom settings. Additionally, there is a scarcity of publicly available engagement datasets that encompass multiple nonverbal cues. By constructing a multi-cues classroom engagement database and proposing a bimodal recognition method, this research seeks to enhance the understanding and measurement of student engagement through a more holistic approach.

**What are the key findings or conclusions of this study?**

The key findings of the study include:

1. **High Accuracy in Engagement Recognition**: The proposed bimodal learning engagement recognition method achieved an accuracy of 93.94% using the KNN classifier, demonstrating that it outperforms many existing state-of-the-art techniques in recognizing student engagement from classroom videos.
2. **Effective Use of Nonverbal Cues**: The study successfully utilized nonverbal cues, such as facial expressions, hand gestures, and body posture, to assess both emotional and behavioral engagement levels. This approach highlights the importance of multiple nonverbal indicators in understanding student engagement.
3. **Dataset Construction**: The research contributed to the field by creating a self-built dataset that includes various engagement levels (low, medium, high) based on non-invasive classroom videos, which is essential for training and validating engagement recognition models.
4. **Future Research Directions**: The study suggests that future work could expand the dataset to include other modalities, such as verbal and physiological data, and explore different fusion methods to enhance engagement recognition further.

Overall, the findings indicate that a bimodal approach to learning engagement recognition can significantly improve the accuracy and effectiveness of monitoring student engagement in educational settings.

 **What methods or techniques were used in this study?**

The study employed several methods and techniques for recognizing learning engagement:

1. **Bimodal Engagement Recognition Method**: The researchers proposed a bimodal approach that combines emotional and behavioral engagement recognition. This method utilizes both facial expressions (emotional cues) and upper body movements (behavioral cues) to assess student engagement levels.
2. **Transfer Learning with ResNet50**: For the emotional engagement channel, the study applied transfer learning using the ResNet50 model as a pre-trained network. This model was fine-tuned to recognize student emotional engagement based on facial expressions.
3. **CoAtNet for Behavioral Engagement**: In the behavioral engagement channel, the researchers trained a self-built behavioral engagement dataset using the CoAtNet network. This network was specifically designed to estimate student behavioral engagement from video data.
4. **Power IoU Loss Function in YOLOv5**: To detect students in the classroom videos, the study introduced a power Intersection over Union (IoU) loss function based on the YOLOv5 object detection framework. This adaptation aimed to improve detection accuracy in the complex and occluded environments typical of real classrooms.
5. **Data Collection and Annotation**: The dataset was created from videos of 28 students during regular lessons, with engagement levels categorized as low, medium, or high based on emotional and behavioral aspects. The data was annotated to facilitate training and evaluation of the engagement recognition models.

These methods collectively contributed to the development of a robust system for automatically recognizing and analyzing student engagement in classroom settings.

 **How was the methodology designed to address the research question?**

The methodology was designed to address the research question of how to effectively recognize and measure student engagement in offline classroom settings through the following key components:

1. **Bimodal Approach**: The methodology incorporated a bimodal recognition system that utilized both emotional and behavioral engagement cues. This dual approach allowed for a more comprehensive assessment of student engagement, acknowledging that both types of cues are essential for understanding engagement levels in a classroom environment , .
2. **Data Collection and Annotation**: The researchers collected a dataset from videos of 28 students during regular lessons, ensuring that the data reflected real classroom dynamics. The videos were annotated to categorize engagement levels into low, medium, and high, based on observable emotional and behavioral indicators. This systematic data collection and annotation process provided a solid foundation for training and validating the engagement recognition models , .
3. **Use of Advanced Deep Learning Models**: The methodology employed advanced deep learning techniques, including transfer learning with ResNet50 for emotional engagement and training a CoAtNet network for behavioral engagement. These models were chosen for their ability to effectively learn from complex visual data, enhancing the accuracy of engagement recognition , .
4. **Object Detection with YOLOv5**: The integration of the YOLOv5 framework with a power IoU loss function allowed for precise detection of students in the videos, even in challenging conditions where occlusion might occur. This step was crucial for ensuring that the engagement recognition models had accurate input data to analyze , .
5. **Classifier Selection and Optimization**: The methodology included the selection of optimal classifiers and parameter tuning to build a general, person-independent engagement model. This approach aimed to prevent overfitting to the training data and ensure that the model could generalize well to new, unseen data , .

By combining these elements, the methodology effectively addressed the research question by providing a robust framework for automatically recognizing and analyzing student engagement in real classroom settings, leveraging both emotional and behavioral cues for a more holistic understanding of engagement.

 **What are the key steps involved in the experimental or analytical process?**

The key steps involved in the experimental or analytical process of the study are as follows:

1. **Dataset Construction**:
   * **Data Collection**: Videos were collected from 28 students (6 male and 22 female) during regular lessons in a smart classroom setting. Consent was obtained from participants for recording their performances.
   * **Video Processing**: A total of 12 videos, each 45 minutes long, were recorded. Frames were extracted from these videos at a rate of one frame every 6 seconds, resulting in approximately 4,550 samples for analysis.
2. **Data Annotation**:
   * The engagement levels were categorized into three classes: low, medium, and high. This categorization was based on emotional and behavioral indicators, such as facial expressions, hand gestures, and body posture.
3. **Model Development**:
   * **Emotional Engagement Recognition**: The ResNet50 model was utilized for recognizing emotional engagement based on facial expressions. Transfer learning techniques were applied to fine-tune the model for this specific task.
   * **Behavioral Engagement Recognition**: A self-built behavioral engagement dataset was used to train the CoAtNet network, which was designed to estimate student behavioral engagement from video data.
4. **Object Detection**:
   * The YOLOv5 framework was employed to detect students in the classroom videos. The power IoU loss function was introduced to enhance detection accuracy, particularly in complex classroom environments where occlusion might occur.
5. **Classifier Training and Evaluation**:
   * Various classifiers, including KNN, were trained on the extracted features from the emotional and behavioral engagement models. The performance of these classifiers was evaluated to determine their accuracy in recognizing student engagement levels.
6. **Results Analysis**:
   * The accuracy of the proposed bimodal engagement recognition method was reported, showing an overall accuracy of 93.94% with the KNN classifier. The results were compared with existing state-of-the-art techniques to validate the effectiveness of the proposed approach.
7. **Discussion and Future Work**:
   * The study concluded with a discussion of the findings, limitations, and potential future research directions, including the exploration of additional modalities for engagement recognition.

These steps collectively contributed to the development and validation of a robust system for recognizing student engagement in real classroom settings, leveraging advanced deep learning techniques and a comprehensive dataset.

 **Were any specific algorithms, models, or frameworks applied? If so, which ones?**

Yes, the study applied several specific algorithms, models, and frameworks as part of its methodology:

1. **ResNet50**: This deep learning model was used for recognizing emotional engagement based on facial expressions. The researchers employed transfer learning with ResNet50, fine-tuning it to adapt to the specific task of emotional engagement recognition.
2. **CoAtNet**: This model was utilized for estimating behavioral engagement. The researchers trained the CoAtNet network using a self-built behavioral engagement dataset, which focused on analyzing nonverbal cues such as body posture and hand gestures.
3. **YOLOv5**: The YOLOv5 framework was implemented for object detection within the classroom videos. The study introduced a power Intersection over Union (IoU) loss function to enhance the detection accuracy of students in complex classroom environments.
4. **K-Nearest Neighbors (KNN)**: This algorithm was one of the classifiers used to evaluate the engagement recognition performance. The KNN classifier achieved an accuracy of 93.94%, demonstrating its effectiveness in classifying bimodal learning engagement.
5. **Other Classifiers**: The study also compared the performance of several other classifiers, including decision trees (DT), random forest (RF), naive Bayes (NB), logistic regression (LR), and support vector machines (SVM). The KNN classifier outperformed these models in terms of accuracy.

These algorithms and models were integral to the study's approach, enabling the researchers to effectively recognize and analyze student engagement through a combination of emotional and behavioral cues.

**Datasets Used in the Study**

1. **Multi-Cues Classroom Learning Engagement Database**:
   * This dataset was specifically constructed for the study and is based on non-invasive classroom videos.

**Dataset Collection and Curation**

* **Data Collection**:
  + The dataset was collected from videos of 28 students (6 male and 22 female) during regular lessons in a smart classroom. The classroom was equipped with multiple cameras positioned to capture the students' interactions and behaviors.
  + Consent was obtained from both the teacher and students for recording their performances in the classroom setting.
  + A total of 12 videos, each 45 minutes long, were recorded in MP4 format. From these videos, frames were extracted at a rate of one frame every 6 seconds, resulting in approximately 4,550 samples for analysis.
* **Data Annotation**:
  + The engagement levels were categorized into three classes: low, medium, and high. This categorization was based on observable emotional and behavioral indicators, such as facial expressions, hand gestures, and body posture.

**Key Characteristics of the Dataset**

* **Size**:
  + The dataset consisted of approximately 12,850 samples labeled as low engagement, 12,100 samples labeled as medium engagement, and 8,380 samples labeled as high engagement, totaling around 33,330 labeled instances.
* **Type**:
  + The dataset included video data that captured various nonverbal cues related to student engagement, such as facial expressions, hand gestures, and body posture.
* **Variables**:
  + The key variables in the dataset included:
    - Emotional engagement indicators (e.g., facial expressions)
    - Behavioral engagement indicators (e.g., hand gestures, body posture)
    - Engagement level labels (low, medium, high)

**Tools and Software Used for Data Analysis or Modeling**

* **Deep Learning Framework**:
  + The study utilized **PyTorch** as the deep learning framework for implementing the models and conducting the analysis.
* **Object Detection Framework**:
  + **YOLOv5** was used for object detection to identify students in the classroom videos , .
* **Model Training and Evaluation**:
  + Various classifiers, including KNN, decision trees, random forest, naive Bayes, logistic regression, and support vector machines, were employed for training and evaluating the engagement recognition models.
* **Computational Environment**:
  + The experiments were conducted on a Windows 10 64-bit system equipped with an Intel(R) Xeon(R) Silver 4112 CPU and an NVIDIA TITAN V GPU, utilizing CUDA for acceleration.

These elements combined to create a robust dataset and analytical framework for studying student engagement in classroom settings.

**Significant Results of the Study**

1. **Accuracy of Engagement Recognition**:
   * The proposed bimodal learning engagement recognition method achieved an accuracy of **93.94%** when using the K-Nearest Neighbors (KNN) classifier. This performance was notably higher than other methods previously reported in the literature , .
2. **Comparison with Other Methods**:
   * The study's method outperformed previous engagement recognition approaches, such as the unobtrusive engagement recognition method by Ashwin et al., which achieved **71% accuracy**. This indicates a significant improvement in recognizing student engagement through nonverbal cues.
3. **Dataset Construction**:
   * The self-built dataset included a substantial number of labeled instances (33,330) across three engagement levels (low, medium, high), which is a notable contribution to the field, as there are few public datasets available for engagement analysis in real classroom settings.

**Comparison to Previous Work**

* The results of this study demonstrate a marked improvement over earlier works that primarily focused on single modalities or had lower accuracy rates. For instance, previous studies often relied on limited datasets or specific nonverbal cues, which restricted their effectiveness in recognizing engagement comprehensively. The integration of multiple nonverbal cues (facial expressions, hand gestures, body posture) in this study allowed for a more holistic approach to engagement recognition.
* The study also highlighted the challenges faced in previous research, such as the complexity and occlusion in real classroom environments, which were effectively addressed through the use of advanced detection techniques like YOLOv5 and the power IoU loss function.

**Surprising Findings or Outliers**

* One surprising finding was the high accuracy achieved with the KNN classifier, especially considering the complexity of the classroom environment and the variability in student behavior. The study's results suggest that even simple classifiers can perform well when combined with a robust dataset and effective feature extraction methods.
* Additionally, the study noted that while emotional and behavioral dimensions were effectively captured, the cognitive dimension of engagement, which is more challenging to identify, was not included in the current analysis. This opens up avenues for future research to explore cognitive engagement and its impact on overall student engagement.

Overall, the study's findings contribute significantly to the field of learning engagement recognition, providing a strong foundation for future research and applications in educational settings.

**Validation of the Proposed Method or Model**

The proposed bimodal learning engagement recognition method was validated through a systematic experimental setup that included the following steps:

1. **Dataset Splitting**:
   * The self-built behavioral engagement dataset was divided into three parts: **17,031 images for training**, **4,866 images for validation**, and **2,433 images for testing**. This split allowed for a robust evaluation of the model's performance on unseen data.
2. **Model Training**:
   * The CoAtNet network was trained using the training dataset, employing softmax cross-entropy loss to predict the engagement levels (low, medium, high). The training process utilized an Adam optimizer to adjust the model parameters effectively.
3. **Testing and Evaluation**:
   * After training, the model was tested on the separate test set to evaluate its performance in recognizing student engagement levels.

**Metrics Used to Evaluate Performance**

The performance of the proposed method was evaluated using several key metrics:

1. **Accuracy**:
   * The primary metric reported was the overall accuracy of the engagement recognition, which was found to be **93.94%** using the KNN classifier. This metric indicates the proportion of correctly classified instances out of the total instances.
2. **Mean Average Precision (mAP)**:
   * The mAP was used to assess the performance of the YOLOv5 model in detecting students within the classroom videos. This metric provides a comprehensive measure of the model's precision and recall across different classes.
3. **Confusion Matrix**:
   * Although not explicitly mentioned in the provided excerpts, confusion matrices are commonly used in classification tasks to visualize the performance of the model across different classes, helping to identify any misclassifications.

**Benchmarks and Comparisons with Existing Methods**

* The study compared its results with existing methods in the field, particularly highlighting the performance of Ashwin et al.'s method, which achieved **71% accuracy**. The significant improvement to **93.94%** accuracy with the proposed method demonstrates its effectiveness in recognizing student engagement through multiple nonverbal cues.
* The study also referenced other works that focused on different aspects of engagement recognition, such as head pose estimation and facial expression recognition, but noted that these methods often lacked the comprehensive approach that combines emotional and behavioral cues as done in this research.
* The introduction of the power IoU loss function in YOLOv5 for student detection was another innovative aspect that contributed to the model's high performance, achieving a precision of **95.4%** in detecting students in complex classroom environments.

Overall, the validation process, combined with the use of robust metrics and comparisons to existing methods, underscores the effectiveness and advancements of the proposed bimodal engagement recognition method in the context of classroom learning engagement analysis.

The study on bimodal learning engagement recognition presents several practical applications and implications, as well as considerations regarding ethics and societal impacts. Here are the responses to your queries based on the content of the PDF:

1. **Practical Applications or Implications**: The findings suggest that recognizing student engagement through nonverbal cues can help educators tailor their teaching strategies to better meet students' needs. By monitoring engagement levels in real-time, teachers can adjust their instructional methods, provide timely interventions, and enhance overall classroom dynamics.
2. **Real-World Scenarios**: The results could be applied in various educational settings, such as traditional classrooms, online learning environments, and hybrid models. For instance, schools could implement video analysis tools to assess student engagement during lessons, allowing for immediate feedback and adjustments to teaching approaches. Additionally, educational technology companies could integrate these recognition methods into their platforms to enhance user experience and learning outcomes.
3. **Ethical Considerations**: While the study does not explicitly detail ethical considerations, it implies the importance of privacy and consent when using video data for engagement analysis. Ensuring that students are aware of and consent to being recorded for educational purposes is crucial. Moreover, the study emphasizes the need for non-invasive methods to respect students' personal space and comfort.
4. **Societal or Environmental Impacts**: The study does not specifically mention societal or environmental impacts. However, the broader implications of improved engagement recognition could lead to enhanced educational outcomes, which may contribute to a more educated workforce and informed society. Additionally, effective engagement strategies could reduce dropout rates and improve overall educational equity.
5. **Foundational References or Related Works**: The study cites several foundational works, including those by Ventura et al. and Ashwin et al., which explore nonverbal cues in engagement analysis. It also references various methodologies in affective computing and engagement tracing, highlighting the evolution of research in this area.
6. **Supplementary Material or Data**: The study mentions the construction of a self-built dataset containing multiple nonverbal cues for engagement analysis. However, it does not specify how to access this dataset or if any supplementary material is provided. Interested researchers may need to contact the authors directly for access to the dataset or additional resources.

Dataset : Not public.