360-Degree Cameras vs Traditional Cameras in Multimodal Learning Analytics: Comparative Study of Facial Recognition and Pose Estimation

Robin Jephthah Rajarathinam
University of Illinois Urbana-Champaign
Champaign, IL, USA
rjrthnm2@illinois.edu

Chris Palaguachi
University of Illinois Urbana-Champaign
Champaign, IL, USA
cwp5@illinois.edu

Jina Kang
University of Illinois Urbana-Champaign
Champaign, IL, USA
jinakang@illinois.edu

Multimodal Learning Analytics (MMLA) has emerged as a powerful approach within the computer-supported collaborative learning community, offering nuanced insights into learning processes through diverse data sources. Despite its potential, the prevalent reliance on traditional instruments such as tripod-mounted digital cameras for video capture often results in suboptimal data quality for facial expressions and poses captured, which is crucial for understanding collaborative dynamics. This study introduces an innovative approach to overcome this limitation by employing 360-degree camera technology to capture students' facial and body features while collaborating in small working groups. A comparative analysis of 1.5 hours of video data from both traditional tripod-mounted digital cameras and 360-degree cameras evaluated the efficacy of these methods in capturing facial action units (AUs) and face and body keypoints. The use of OpenFace revealed that the 360-degree camera captured high-quality facial features more effectively than the traditional method, significantly enhancing the reliability of facial feature detection. Similarly, OpenPose analysis demonstrated that the 360-degree camera substantially improved the capture of complete body keypoints compared to the traditional setup. These findings suggest that integrating 360-degree camera technology in MMLA can provide richer data for analyzing affect and engagement in collaborative learning environments. Future research will focus on refining this technology to further enhance our understanding of the learning process.

Keywords: facial features, pose estimation, engagement, affect detection, computer-supported collaborative learning, multimodal learning analytics, video recording

1. INTRODUCTION

In recent years, the field of education has witnessed a transformative shift towards leveraging advanced technologies to enhance our understanding of learning. Among these innovations,

Multimodal Learning Analytics (MMLA) has emerged as a pivotal approach, offering nuanced insights into the dynamics of collaborative learning. MMLA expands the scope of traditional learning analytics by integrating a diverse array of data sources, encompassing not only digital interactions but also leveraging sophisticated sensory technologies. This approach facilitates a deeper understanding of the complex interplay between cognitive, affective, and engagement factors in learning environments. Despite the significant advancements in MMLA, the practical deployment of these technologies in real-world educational settings presents a myriad of challenges, ranging from technical hurdles to the intricacies of data interpretation (Martinez-Maldonado et al., 2023). Moreover, while the affective and engagement dimensions of learning have gained increasing recognition for their impact on educational outcomes, the methodologies for capturing high quality affective and engagement features in collaborative settings remain underexplored.

A critical challenge in MMLA is the effective capture of frontal or near-frontal views of students collaborating around a table, which is essential for accurate analysis of facial expressions and body movement. Traditional side-located cameras often fail to capture these views due to obstructions and limited angles, resulting in suboptimal data quality. Previous studies have attempted to address this issue using multiple cameras placed at the center of the table (Noël et al., 2022) or employing fisheye cameras (Lewis et al., 2023). While these approaches have shown promise, they introduce complexities such as increased equipment, potential obstructions, and image distortion that complicate data processing.

This study introduces the use of 360-degree camera technology as an alternative solution to overcome these limitations. By capturing comprehensive, uninterrupted views of all participants from a single device, 360-degree cameras reduce setup complexity and minimize obstructions in the collaborative space. Additionally, our exploratory research evaluates the efficacy of 360-degree camera technology in enhancing video data quality for face-to-face collaborative learning scenarios. By tackling these challenges, the study seeks to contribute to the ongoing discourse on improving the accuracy and reliability of multimodal data analysis in educational research through innovative data collection and processing methods.

This paper builds on our previous conference paper (Rajarathinam et al., 2024) and additionally presents pose estimation results not included in the earlier work. We also add a more comprehensive discussion and visualization of results to investigate the impact of 360-degree cameras on facial and body features estimation.

The remainder of this article is organized as follows. First, we will review the related work. Then, we introduce our task and present the framework in detail. Next, we introduce the context and the dataset in Section 3 and report the results in Section 4. Finally, we discuss the results and the limitations of our work in Section 5 and conclude the paper in Section 6.

2. Background

2.1. MMLA in Collaborative Learning

Collaborative learning is a complex sense-making process in which a group of students works together to co-construct knowledge via iterative social interactions (Roschelle, 1992). During group collaboration, students present their ideas and explain to their peers how to understand concepts, solve tasks, and justify ideas in response to questions, challenges, and conflicts. MMLA represents a significant evolution in offering a sophisticated framework for understand-

ing the intricacies of group learning dynamics in collaborative learning environments. By extending beyond the traditional analysis of student interactions through digital platforms, which predominantly utilize input devices like keyboards and mice, MMLA captures the richness of collaborative interactions. These interactions are characterized by varying degrees of technological mediation, from fully (Vrzakova et al., 2020) to partly (Schneider et al., 2018), or even completely unmediated scenarios (Sümer et al., 2023), encompassing interactions not only between students and teachers (D'Mello et al., 2015) but also within the tangible learning environments (Yan et al., 2021). A focal point of MMLA in collaborative settings is the exploration of learner attributes that are difficult to quantify without advanced sensing technologies. This includes analyzing emotional states (Ma et al., 2024), cognitive conditions (Prata et al., 2009), distractions (Liao and Wu, 2022), and stress levels (Ronda-Carracao et al., 2021) within group contexts. Initially, video and audio recordings served as the principal data sources for MMLA due to the limitations of available technology. As sensing technologies have advanced, MMLA has seen a marked evolution, incorporating a broader range of data modalities. The practical application of MMLA in such environments often involves the use of diverse sensors, including eye-trackers, positioning systems, wearable microphones, and physiological sensors on wrists or chests, alongside sophisticated audio and video processing algorithms (Alwahaby et al., 2022). This technology suite generates a wealth of multimodal data, enabling a comprehensive analysis of the complex phenomena inherent in collaborative learning.

2.2. Affect Detection

Learners may experience a variety of cognitive-affective states when they are assigned difficult tasks to solve, including confusion, frustration, boredom, engagement/flow, curiosity, anxiety, delight, and surprise (Graesser and D'Mello, 2012). These experiences underscore the complexity of the learning process, where cognitive and affective dimensions intertwine, necessitating a comprehensive approach to understanding learning dynamics. Building upon cognitive foundations, the realm of MMLA equally delves into the affective dimensions of learning. Central to this exploration is the Control-Value Theory of Achievement Emotions (CVTAE; Pekrun 2006), a cornerstone in affective domain research within MMLA. CVTAE's application across MMLA studies stems from its pivotal role in bridging the gap between affective computing and intelligent tutoring systems, highlighting the intricate relationship between learners' affective states and their learning experiences. The integration of CVTAE within MMLA is further supported by advanced analytical frameworks and tools, notably the Facial Action Coding System (FACS) developed by Ekman et al. (2002) and operationalized through technologies like the OpenFace library (Amos et al., 2016). In the context of MMLA, video data emerges as a critical medium for capturing non-verbal cues that are essential for understanding collaborative learning dynamics. Video data allows for the analysis of student-student behaviors that are crucial for group success and individual contributions that facilitate self-assessment and personal growth within group settings (Zhou and Cukurova, 2023). The sophisticated analysis of video data, including facial expressions and body movement, offers insights into the multifaceted affective factors that influence collaborative learning quality.

2.3. Engagement detection

Various approaches have emerged to focus on detecting student engagement, particularly through the analysis of physical body posture and associated behaviors (Baker et al., 2012; Botelho et al.,

2017; Fahid et al., 2023). Learning engagement, defined as the extent of learners' devotion to a learning activity that results in a desired learning outcome (Hu and Kuh, 2002), is positively related to learning achievement (Li and Baker, 2018) and can significantly predict learning success (Liu et al., 2022). Given its crucial role in improving learning outcomes, accurately detecting and understanding engagement is essential. Utilizing computer visual tools (Hur and Bosch, 2022), these methods involve detecting subtle student movements and behaviors, such as leaning towards peers during group discussions, leaning back, crossing arms, or engaging in sharing and pointing to group materials (Radu et al., 2020). Such behaviors provide valuable insights into learners' internal states, including their level of engagement and comfort with collaborators. In addition, these behaviors play a crucial role in informing teachers when to intervene during small group work, based on students' level of engagement in collaborative activities (Taylor, 2016; Radu et al., 2020; Stewart et al., 2021; Hur et al., 2020).

Several tools have been developed to track students' body positions during paired and small group work. For instance, Radu et al. (2020) utilized unsupervised machine learning methods to analyze positional data collected from Microsoft's Kinect V2. They identified several collaborative behaviors, such as turn taking, open to collaboration, closed to collaboration, and synchronized leaning. These behaviors were identified by extracting key positional data points from students (keypoints) and analyzing features such as spine similarity, distance between the participants, eye gaze, and head movements. Another popular method involves algorithms that detect posture keypoints from images. One such toolkit is called OpenPose, a real-time multi-person system that detects the human body, hands, facial, and foot keypoints on single images (Cao et al., 2021). Depending on the model, OpenPose can detect more than 135 keypoints per person, including 25 for the full-body, 2x21 for the hands, and up to 70 for the face. OpenPose utilizes Convolutional Pose Machines for accurately predicting human body keypoints and Part Affinity Fields to connect these keypoints into coherent poses by encoding the spatial relationships between them. By analyzing the positions and movements of these keypoints, researchers can extract detailed information about how individuals or groups physically interact during collaborative tasks, allowing for a more precise analysis of physical engagement and interaction patterns.

However, despite these advancements, tracking multi-person full-body positional data in naturalistic classroom settings continues to pose challenges, particularly in accurately capturing individual students' positional data during small group work. Challenges such as various types of occlusion—caused by other students, teachers, furniture, laptops, or even self-occlusion—along with complex body configurations, diverse appearances including clothing, and the complexities of the environment, such as varying viewing angles, distance from the camera, and truncation of parts in the camera view, can significantly impede accurate data collection (Mishra et al., 2024). Consequently, there is a pressing need for novel approaches to mitigate these challenges and to assess the accuracy of pose estimation tools in tracking students' engagement during small group activities in naturalistic classroom environments.

2.4. Real Classroom Implementation

Traditional camera setups for MMLA often rely on single-camera configurations paired with machine learning algorithms to analyze student interactions (Chejara et al., 2023; Zhou et al., 2023). In controlled lab environments, these approaches have yielded promising results for tracking specific behavioral features. For example, Andrade (2017) employed front-facing we-

bcams to track eye gaze and hand positioning, examining relationships to embodied cognition and learning performance. Although the study showed that such features could serve as indicators of engagement, its small sample size and controlled setting limited the generalizability to real-world classrooms.

Other researchers have used multi-camera setups in controlled contexts to capture richer behavioral data. Spikol et al. (2017) combined front-facing and top-down cameras to measure metrics such as 'Faces Looking at the Screen,' 'Distance Between Hands,' and 'Hand Motion Speed' during collaborative tasks. These features quantified attention and physical proximity among group members but encountered significant occlusion issues, with 55–65% blockage in face and hand tracking. Similarly, Oviatt et al. (2021) deployed five synchronized cameras, including close-up, wide-angle, and top-down views, to track hand gestures and group interactions. Their findings revealed that increased use of iconic gestures during complex tasks associated with higher collaboration quality. However, manual pre-processing demands and the limited granularity of tools like OpenPose constrained scalability.

When MMLA tools transition from controlled studies to naturalistic classroom settings, additional challenges emerge. Variability in camera placement, physical constraints, and ambient noise can degrade data quality. For instance, Dai et al. (2023) applied OpenFace to teacher-recorded classroom videos to detect negative facial expressions. Partially visible faces due to suboptimal camera angles led to missed detections and reduced reliability. Similarly, Chejara et al. (2023) used CoTrack and Etherpad to analyze facial action units, head poses, and speech overlap, successfully highlighting the importance of expressions and head orientation for gauging engagement. Yet, the use of front-facing webcams alone limited the capture of other crucial features, such as body positioning and hand movements.

To capture broader classroom dynamics, some researchers have built or customized devices. Lewis et al. (2023) integrated a wide-angle lens (220-degree field of view) with a 4-microphone array, placing the system at the center of each group. This wider vantage point captured seating arrangements, proximal gestures, and speech patterns across 79 students, offering richer insights into collaboration. Nevertheless, occlusion from face masks (as this study was conducted during COVID-19 precautions) and limited classroom space still led to degraded data quality.

These studies show that traditional and multi-camera approaches can provide detailed information on student behaviors. However, they also reveal significant limitations in crowded or dynamic classroom environments. Building on this prior work, our study offers practical guidance for optimizing data collection and improving reliability when deploying MMLA tools in authentic classroom settings. As we address the benefits and limitations of various camera configurations, it is crucial to remain aware of persistent challenges. These challenges range from classroom size and device setup to noise and limited algorithmic robustness for facial and body detection.

2.5. 360-DEGREE CAMERA IN EDUCATION

The advent of 360-degree camera technology introduces a promising avenue for capturing comprehensive classroom footage from a single vantage point. Unlike traditional or multi-camera setups that require multiple devices and careful synchronization, 360-degree cameras use omnidirectional or multi-lens systems to record footage from every angle simultaneously and stitch it into a spherical view. This approach enables viewers to explore the captured environment in any direction, offering a panoramic perspective that can facilitate richer analyses of teaching and

learning (Mallavarapu et al., 2022; Evens et al., 2023; Noël et al., 2022).

Researchers have leveraged such cameras to create immersive learning experiences (Tedre et al., 2021), develop computer vision programs for robotics, and study collaborative learning. Once recorded, 360-degree video can be accessed via various devices—from smartphones and computers (non-immersive) to head-mounted displays (immersive)—allowing flexibility in how data is reviewed (Snelson and Hsu, 2020). This versatility also enables diverse perspectives,including student facial views and partial top-down views for hand tracking (Oviatt et al., 2021; Spikol et al., 2017).

Despite these benefits, the application of 360-degree cameras within MMLA remains limited. For example, Noël et al. (2022) used machine learning techniques (Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN)) with multidirectional microphones and USB cameras placed at the center of a small group to analyze collaborative behaviors, including spoken interaction distribution and joint gestures. Although the study successfully detected cognitive, metacognitive, and affective learning behaviors, scaling this approach to multiple groups in a typical classroom remains challenging due to the number of cameras required and the associated processing demands.

To address this gap, our study leverages a 360-degree camera for capturing richer, uninterrupted views of small-group collaboration. Positioned at the group's center, this setup minimizes complexity and potentially reduces occlusion issues. Although 360-degree cameras require maintaining clear space around the camera to avoid distortion, this minor limitation is offset by their ability to provide a more holistic and scalable approach to analyzing face-to-face collaborative behaviors in realistic classroom environments. Integrating this spherical video capture into MMLA research thus represents an important methodological step forward, allowing for deeper insights into the complexities of collaborative learning in authentic educational contexts.

2.6. Research Questions

To assess the effectiveness of this technological intervention, the research is guided by three critical research questions (RQs):

(RQ 1) How does the effectiveness of utilizing a 360-degree camera compare to traditional video capture methods in detecting facial and body features within small group interactions in the classroom?

(RQ 2) How effectively does OpenFace, a facial recognition toolkit, extract facial features from all students simultaneously during small group work in the classroom when using 360-degree cameras versus traditional cameras?

(RQ 3) How effectively does OpenPose, a posture estimation toolkit, extract body features from all students simultaneously during small group work in the classroom when using 360-degree cameras versus traditional cameras?

3. Methods

3.1. Context

The dataset comprises classroom observations from a group activity during the discussion session of an introductory digital learning environment course at a Midwestern University. The

objective of the activity was to engage students with a web-based immersive science learning environment, HoloOrbits (Figure 1). HoloOrbits was initially developed for the Microsoft HoloLens2 on Universal Windows Platform within Unity. A new version of HoloOrbits was developed for laptops using WebGL via Unity to make the simulation more accessible to students in classrooms. The direct goal of the simulation is to help students learn about planetary motion and Kepler's laws. The simulation immerses students in the factual and conceptual understanding of the elliptical orbits within a newly discovered exoplanetary system. HoloOrbits offers tools enabling students to simulate abstract components of the planetary system and collect data (e.g., distances between celestial objects) necessary for understanding the orbital system and Kepler's laws. The main learning goal is to create experiences that support students in grasping scientific concepts and foster agency by empowering them to conduct their own scientific investigations. Following this main activity, the students worked in groups to reflect on the task and the design of the simulation. The entire activity lasted approximately 1 hour and 30 minutes, with the students taking a 10-minute break between the interaction and reflection phases of the activity.

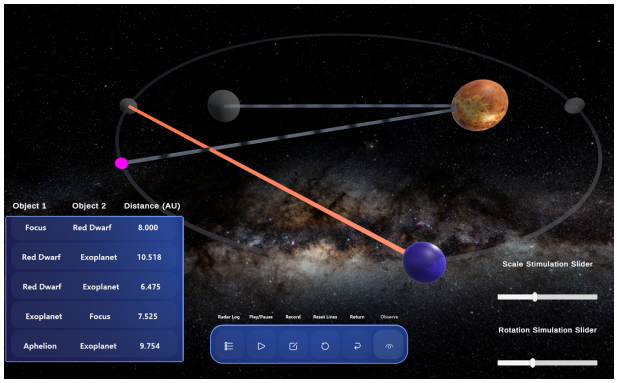

Figure 1: HoloOrbits: A web-based immersive science learning simulation

3.2. Participants

The dataset collected consisted of two different classroom observations. A total of 24 out of the 39 students consented to participate in the research. The participants worked in groups of 2 to 4 students. Ethical approval was obtained from the institutional review board of the authors' institution. We collected the demographics of the participants. Local devices were used for storing and processing participants' data. We used open-source tools, de-identified the data, and used pseudonyms after processing it via OpenFace. This preliminary study focused on two groups within the same classroom observation. One group contained four students whose video was captured using a GoPro 360-degree Max camera (Yellow Group) and another group with three students whose video was captured using a tripod and (traditional, non-360) GoPro Hero10 camera (Red Group) (see Figure 2). These two groups were ideal as each student had their own devices to work on the stimulation. Both groups were positioned in the corners of the classroom, however, the red group benefited from a more controlled viewing angle, as the traditional camera was pointed towards a corner of the classroom (out of the view of other students).

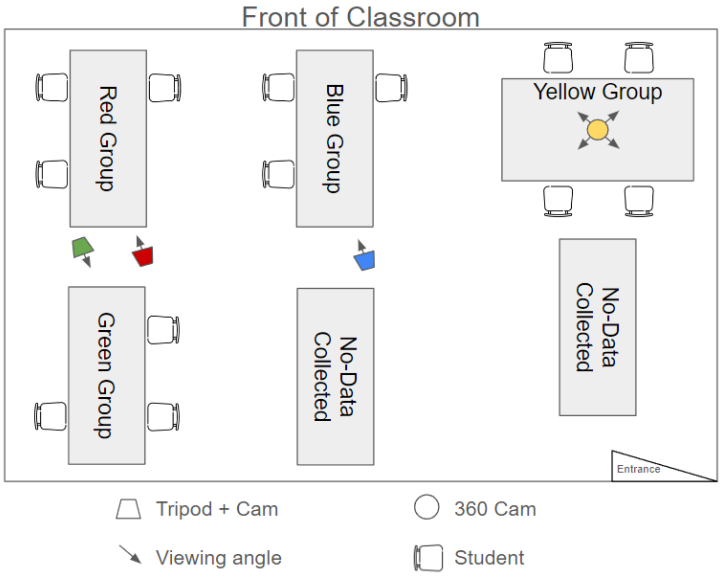

Figure 2: Schematic of the classroom for data collection

3.3. PRE-PROCESSING

The 360-degree camera generated .360 files, which were converted using .H264 encoder with equilateral projection into .mp4 files (ProRES files) using GoPro's proprietary software (GoPro Player). These videos were stitched together as GoPro capture video in chunks of 20 mins. Adobe Premiere Pro v24.1 was used to convert ProRES .mp4 files to create the Top-Down view (TDV) as well as Face view (FV) (see Figure 3). FV was setup as a means to isolate and maximize the exposure of individuals' faces to the camera lens, while TDV was set up as a means to isolate and maximize the exposure of the groups' physical interactivity to the camera lens. Both views were intended to capture peer-to-peer collaborative interactions. We used an Adobe plugin for GoPro 360-degree footage to modify the direction of the cameras to get both views. For the tripod setup, the output .mp4 files were stitched using Adobe Premiere v24.1 to obtain the traditional view (TV). The final videos were 24 fps with 1080p resolution in .mp4 format. The students were tagged as S1, S2, S3, and S4 starting from the top-left going clockwise for both FV and TDV.

OpenFace v2.2.0 facial behavior analysis toolkit (Baltrusaitis et al., 2018) was used to extract facial features. The multiple faces mode was used to extract the following three categories of facial features: (1) eye gaze direction, (2) head pose, and (3) facial action units (AUs) (see Figure 4). OpenFace v2.2.0 provides confidence of the predicted values for each frame which range from 0 to 1. Any frame with confidence greater than or equal to 0.75 has a success variable (binary) equal to 1. These frames contain facial action units (AUs) which are instrumental in detecting affect of students who are being detected in the group.

OpenPose v1.7.0 pose estimation toolkit (Cao et al., 2021) was used to extract body position features (See Figure 5). The body25 model was used to extract 25 body keypoints. In our analysis, we focused on the 13 upper body keypoints, specifically the Nose (0), Neck (1), Right Shoulder (2), Right Elbow (3), Right Wrist (4), Left Shoulder (5), Left Elbow (6), Left Wrist (7), and Mid Hip (8) because the students were mostly seated in their chairs and interactions

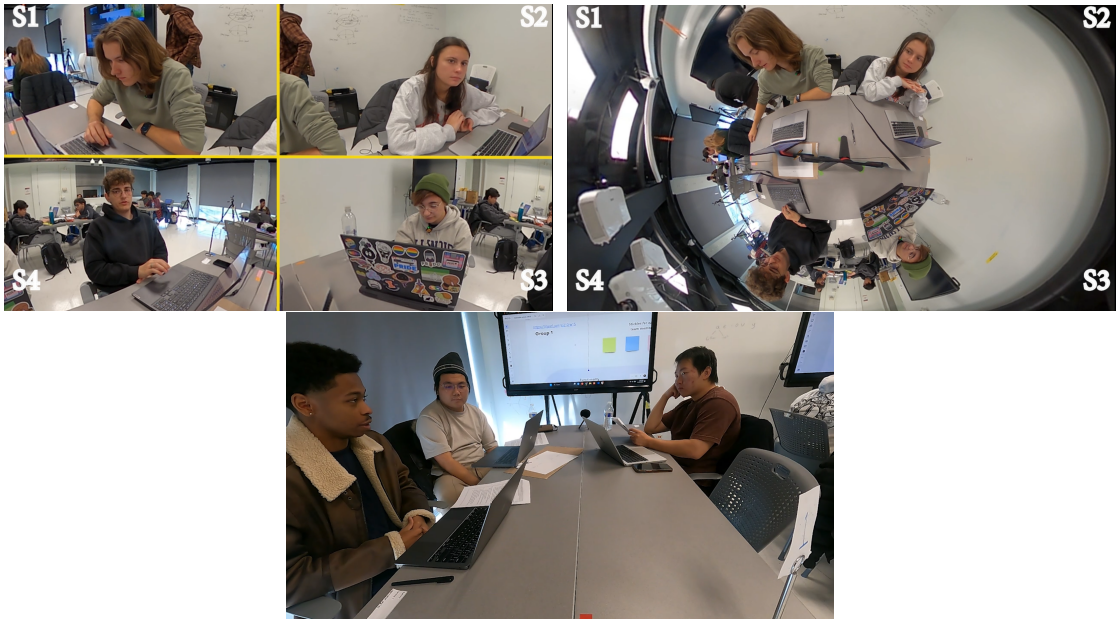

Figure 3: Video capture views: Face View (FV, top-left), Top-Down view (TDV, top-right), and Traditional View (TV, bottom). FV and TDV are from a 360-degree camera, while TV is from tripod setup.

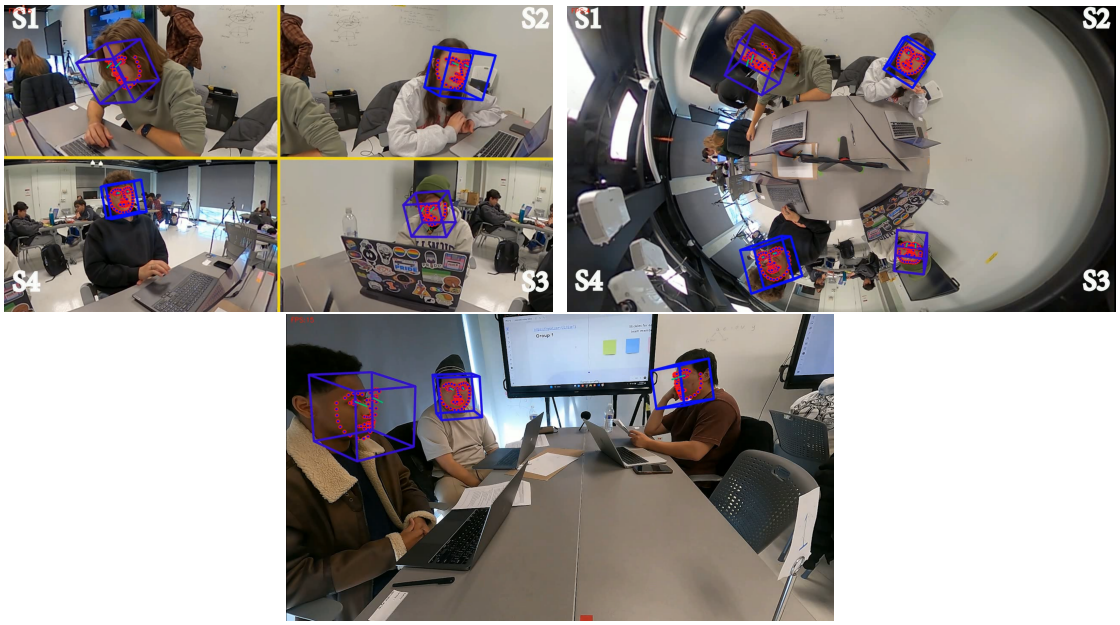

Figure 4: OpenFace facial feature capture: Face View (FV, top-left), Top-Down View (TDV, top-right), and Traditional View (TV, bottom). FV and TDV are from a 360-degree camera, and TV is from a tripod setup.

were using their hands. Additionally, we also considered the Right Eye (15), Left Eye (16), Right Ear (17), and Left Ear (18) to measure head pose. OpenPose v1.7.0 provides confidence scores for the predicted keypoints' coordinates (x, y) in each frame, where 'x' represents the horizontal position and 'y' represents the vertical position within the frame, with scores ranging from 0 to 1. Any frame with a confidence score greater than 0 was considered to have an initial detected keypoint. We then applied a tracking algorithm that assigned ids to students in the video and then manually coded the students of interest. Through this process, we identified that a confidence score of 0.2 or higher was optimal for accurately tracking keypoints (Hur and Bosch, 2022). It is worth noting that this threshold differs from the confidence score threshold used for OpenFace (0.75). The rationale for this difference, including the implications for facial recognition and pose estimation accuracy, is further discussed in Section 3.4.1.

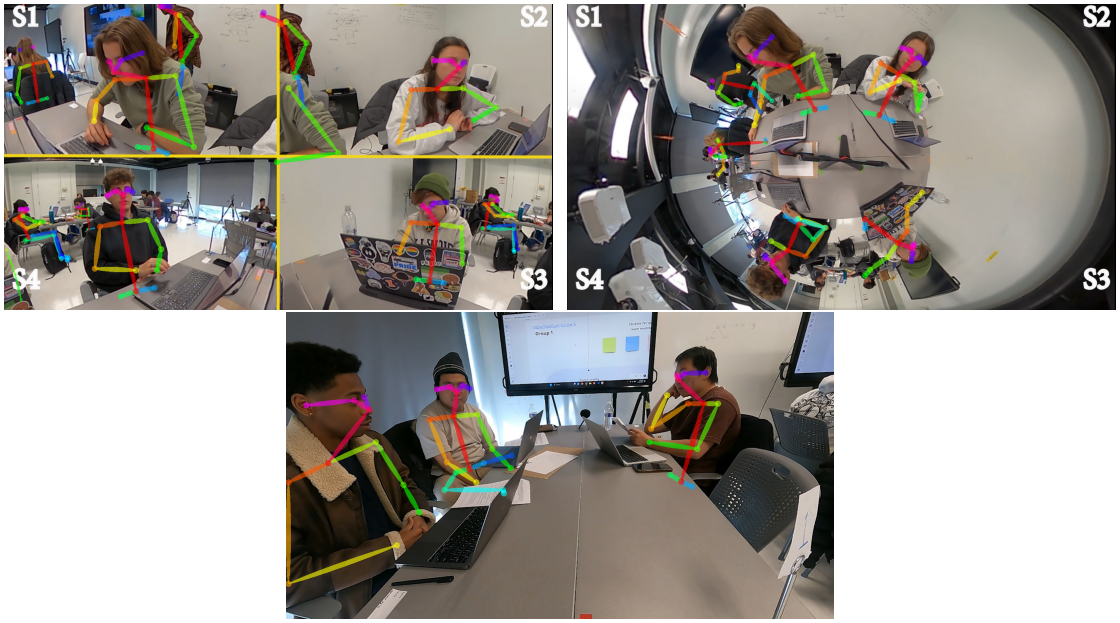

Figure 5: OpenPose body keypoint estimation: Face View (FV, top-left), Top-Down View (TDV, top-right), and Traditional View (TV, bottom). FV and TDV are from a 360-degree camera, while TV is from a tripod setup.

3.4. Analysis

In this study, we investigated the effectiveness of 3 different camera viewing angles in the classroom. To capture the dynamic interactions within small groups, 360-degree cameras were employed to derive two specialized views: a Top-Down View (TDV) and a Face View (FV), designed to provide comprehensive angles for facial feature and body keypoint estimation. Additionally, a Traditional View (TV) was obtained from a conventional video capture setup using a tripod and GoPro Hero10. Additionally, we used a single session to remove external variables like session duration, intervention types, and instructor-student interaction.

To address Research Question 1 (RQ 1), the relative efficacy of 360-degree cameras is compared against traditional video capture methods in detecting facial action units and body position keypoints, aiming to understand their effectiveness in a classroom setting. This is done by conducting statistical analysis using a non-parametric test, the Kruskal-Wallis test, followed by

Dunn's post-hoc test with Bonferroni correction.

To address Research Question 2 (RQ 2), the generation of time series data for TDV, FV, and TV, alongside statistical analysis for high-quality facial feature capture (described in the section 3.4.1) or the three views, enables a rigorous evaluation of OpenFace—a facial recognition toolkit—in accurately extracting facial features from all students simultaneously within small group interactions.

To address Research Question 3 (RQ 3), we generated time series data for the three views: TDV, FV, and TV, alongside statistical analysis to assess high completeness body keypoint estimation (described in the section 3.4.1) in each view. This analysis focused on evaluating OpenPose—a body posture estimation toolkit—in its ability to detect a complete set of upper body keypoints from all students simultaneously during small group work.

R version 4.1 was utilized for comprehensive statistical analysis, encompassing both descriptive statistics and inferential tests such as the Kruskal-Wallis test for non-normally distributed data, followed by Dunn's post-hoc test with Bonferroni correction.

3.4.1. Evaluation criteria

To compare the effectiveness of detection by both OpenFace and OpenPose, we described two evaluation criteria. These criteria serve as metrics to compare the different views being analyzed.

OpenFace evaluation criteria

To evaluate OpenFace's performance on all three views, a Group-wide High quality facial feature detection criteria was established based on two conditions: (1) when OpenFace has a confidence level of 0.75 or higher for each frame captured (Success = 1), and (2) all students in the group are simultaneously detected. Frames that adhere to this criteria will be referred to as *high quality frames* in this paper. A confidence threshold of 0.75 was set as a prerequisite for detecting facial action units by OpenFace. This stringent selection ensures the reliability of future affect detection in collaborative learning environments, highlighting the significance of AUs captured for all group members.

OpenPose evaluation criteria

To evaluate OpenPose's performance on all three views, a Group-wide High completeness detection criteria was established based on 3 conditions: (1) OpenPose must detect at least 10 keypoints (keypoints 0-8, and at least 1 keypoint from 15-18) with a confidence score of 0.2 or higher, and (2) all students in the group must be simultaneously detected. Frames that adhere to these criteria will be later referred to as *high completeness frames* in this paper. Keypoints 0-8 were selected because they identify important positional data such as chest, shoulders, midbody, and arms posture. A more relaxed criteria was selected for a keypoint between 15-18 (left and right ear; left and right eye) because any one of those keypoints would be sufficient enough to determine head axial rotation. Lastly, the confidence threshold of 0.2 was set as a minimum requirement based on the accuracy of keypoint detection from our manual annotation process for student tracking, which aligns with findings in prior research(Hur and Bosch, 2022).

4. RESULTS

4.1. RESEARCH QUESTION 1

How does the effectiveness of utilizing a 360-degree camera compare to traditional video capture methods in detecting facial and body features within small group interactions in the classroom?

Table 1 presents the descriptive statistics of video data from the traditional view (TV) and the two 360-degree camera views (FV and TDV), along with the frames detected using OpenFace and OpenPose. It is important to note that the TV data was collected from one group, while the FV and TDV data were collected from a different group. This distinction introduces potential variability in factors such as group size and individual facial recognition differences, which should be considered when interpreting results. Detection using OpenFace in each view indicates that a face was tracked with a confidence value greater than zero. Similarly, for OpenPose, detection indicates that a body keypoint was tracked with confidence greater than zero. The results revealed significant variations in the efficiency of face and body keypoint capture. The TV and FV angles demonstrated superior detection efficiency with OpenFace, with unique frame counts exceeding those of the TDV angle, suggesting potential detection limitations. In contrast, OpenPose showed much better detection efficiency across all three views for body keypoints. In Table 1, "unique frame detected" refers to any frame where at least one student was successfully tracked by the detection tool (either through face detection by OpenFace or body keypoint detection by OpenPose). This indicates that even if only one student out of the entire group is detected in a given frame, that frame is counted as a unique detection.

Table 1: Descriptive statistics of video frames and frames detected by OpenPose and OpenFace by views

View Angles	Total Frames	Unique Frames (OpenFace)	Unique Frames Percentage (%) (OpenFace)	Unique Frames (OpenPose)	Unique Frames Percentage (%) (OpenPose)
TV	120,504	115,509	95.85	118863	98.64
FV	117,648	116,535	99.05	117482	99.86
TDV	117,648	106,266	90.31	117433	99.82

TV - Traditional View, FV - Face View, TDV - Top-Down View.

The histograms in Figure 6 revealed a non-normal distribution of the confidence levels captured by OpenFace for the frames across the three different viewing angles. For this reason, non-parametric tests were used for testing significant differences between the view angles. The visualizations revealed the differences descriptively, but to confirm the differences statistically, the Kruskal-Wallis test and the Dunn's post-hoc analysis with Bonferroni correction were used.

Figure 7 also revealed a non-normal distribution of the completeness levels of keypoints captured by OpenPose for all the frames across the three different viewing angles. Thus, we used the non-parametric tests, including the Kruskal-Wallis test and Dunn's post-hoc analysis with Bonferroni correction, to determine the differences statistically.

The Kruskal-Wallis test on OpenFace data indicated that there was significant difference in the median confidence values across the three camera views (FV, TDV, and TV) (χ^2 =

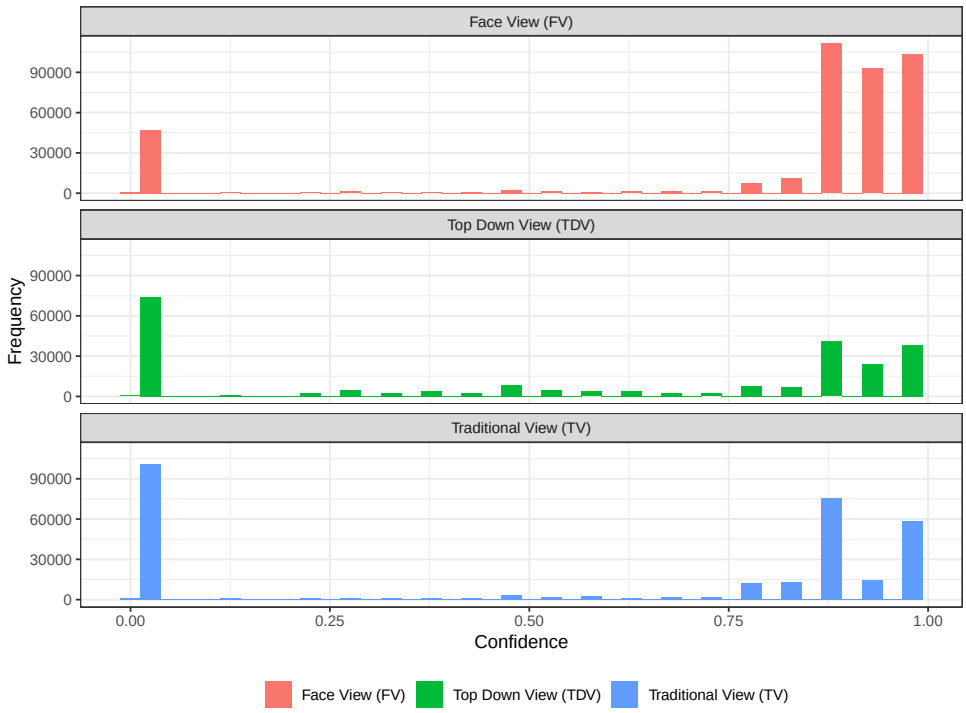

Figure 6: Histograms displaying the distribution of OpenFace confidence levels with frame frequency across three views

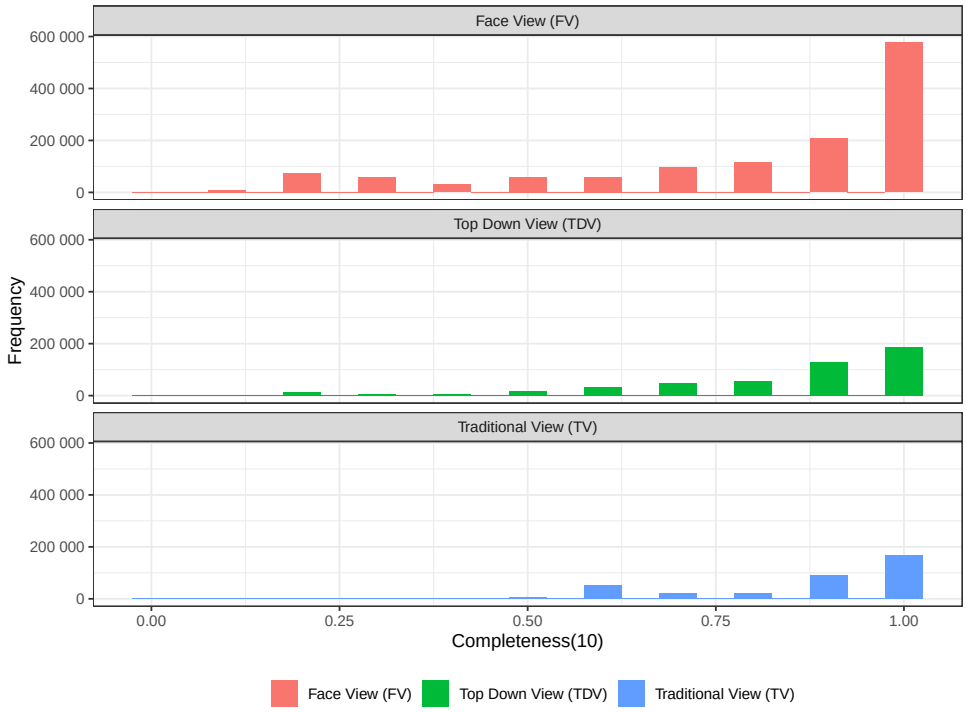

Figure 7: Histograms displaying the distribution of OpenPose completeness levels with frame frequency across three views

82273.378, df = 2, p = 0) with a small to medium effect size(η^2) of 0.054 (For η^2, small effect $\approx$ 0.01; Medium effect $\approx$ 0.06). The Dunn's post-hoc analysis with Bonferroni correction, as shown in Table 2, confirms that each pair of camera views significantly differs in median confidence values. The negative Z-value for the TDV-TV comparison suggests that the rank sum (and thus the median confidence) for TDV is lower than for TV, while positive Z-values for the FV-TDV and FV-TV comparisons suggest higher rank sums (and higher median confidence) for FV compared to TDV and TV, respectively. The significance across all comparisons suggests that the camera views have statistically significant impacts on the confidence values, with FV potentially having the highest median confidence values, followed by TV and then TDV, based on the direction indicated by the Z-values.

Table 2: Summary of Kruskal-Wallis and Dunn's Post-hoc test results for detecting unique frames using OpenFace for the three views

Pairwise Comparison	***Z*-value (OpenFace)**	***p*-value (OpenFace)**	**Adjusted *p*-value (OpenFace)**	**Cohen's r (OpenFace)**
FV vs. TDV	242.35	$<$ .0001*	$<$ .0001*	0.25
FV vs. TV	238.34	$<$ .0001*	$<$ .0001*	0.25
TDV vs. TV	-18.99	$<$ .0001*	$<$ .0001*	-0.02

* indicates statistical significance.
TV - Traditional View, FV - Face View, TDV - Top-Down View.
Cohen's r: Small effect: $r \approx 0.1$, Medium effect: $r \approx 0.3$; Large effect: $r \approx 0.5$

Another Kruskal-Wallis test on OpenPose data indicated that there was significant difference in the median completeness values across the three camera views (FV, TDV, and TV) (χ^2 = 63644.54, df = 2, p = 0) with a medium to large effect size (η^2) of 0.089 (For η^2, medium effect $\approx$ 0.06; large effect $\approx$ 0.14). The Dunn's post-hoc analysis with Bonferroni correction, as shown in Table 3, further supports these findings by revealing significant differences between each pair of the camera views. The positive Z-values for the FV-TDV and FV-TV comparisons suggest that the rank sum, and consequently the median completeness, is higher for FV compared to both TDV and TV. Conversely, the negative Z-value for the TDV-TV comparison indicates that the rank sum (and median completeness) for TDV is lower than for TV. These results suggest that, similar to the OpenFace analysis, the camera views significantly influence the completeness values measured by OpenPose, with FV likely yielding the highest median completeness values, followed by TV, and TDV showing the lowest median completeness values, as inferred from the direction of the Z-values.

4.2. Research Question 2

How effectively does OpenFace, a facial recognition toolkit, extract facial features from all students simultaneously during small group work in the classroom when using 360-degree cameras versus traditional cameras?

Table 4 shows a notable disparity in the detection of high-quality frames detected using OpenFace among the three view angles: TV, TDV, and FV. The FV demonstrates a significantly higher efficacy in yielding high-quality frames, with 39,026 frames, which constitutes 33.17% of the total frames detected. Conversely, the TDV angle shows a markedly lower success rate,

Table 3: Summary of Kruskal-Wallis and Dunn's Post-hoc test results for detecting unique frames using OpenPose for the three views

Pairwise Comparison	***Z*-value (OpenPose)**	***p*-value (OpenPose)**	**Adjusted *p*-value (OpenPose)**	**Cohen's r (OpenPose)**
FV vs. TDV	246.03	$< .0001$*	$< .0001$*	0.23
FV vs. TV	155.14	$< .0001$*	$< .0001$*	0.14
TDV vs. TV	-83.79	$< .0001$*	$< .0001$*	-0.08

* indicates statistical significance.
TV - Traditional View, FV - Face View, TDV - Top-Down View.
Cohen's r: Small effect: $r \approx 0.1$, Medium effect: $r \approx 0.3$; Large effect: $r \approx 0.5$

with only 533 high-quality frames detected, amounting to a mere 0.45% of the total frames. The TV angle, while better than TDV, still yields a relatively low number of high-quality frames at 6,215, representing 5.16% of the total frames.

Table 4: Number of frames with high quality facial feature detection and their percentage of total frames

View angles	**High quality frames (OpenFace)**	**Percentage frames (%) (OpenFace)**	**High completeness frames (OpenPose)**	**Percentage frames (%) (OpenPose)**
TV	6,215	5.16	219	0.18
FV	39,026	33.17	16,762	14.25
TDV	533	0.45	474	0.40

TV - Traditional View, FV - Face View, TDV - Top-Down View.

The time series plots in Figure 8 showcase the capture of high quality facial features across the duration of the class session by OpenFace for the three views. The visualizations revealed that high quality facial features were captured for FV at a relatively consistently rate across time when compared to the other views.

Figure 9 includes time series plots, illustrating the detection of facial features over time for each student from two views: FV and TDV, both of which were captured using 360-degree footage. However, there are noticeable differences in performance between FV and TDV. For example, for student 1 (S1), FV S1 and TDV S1 showcase differences in the amount of facial features of the frames captured. These facial features have a confidence level of greater than 0.75. FV captured more frames with facial features than TDV. Particularly, within TDV, facial features captured for S3 and S4 are much worse when compared to S1 and S2. This discrepancy is likely due to the inverted orientation of S3 and S4 in the 360-degree footage, which affected the accuracy of OpenFace's detection. Further elaboration on this challenge is provided in the discussion section.

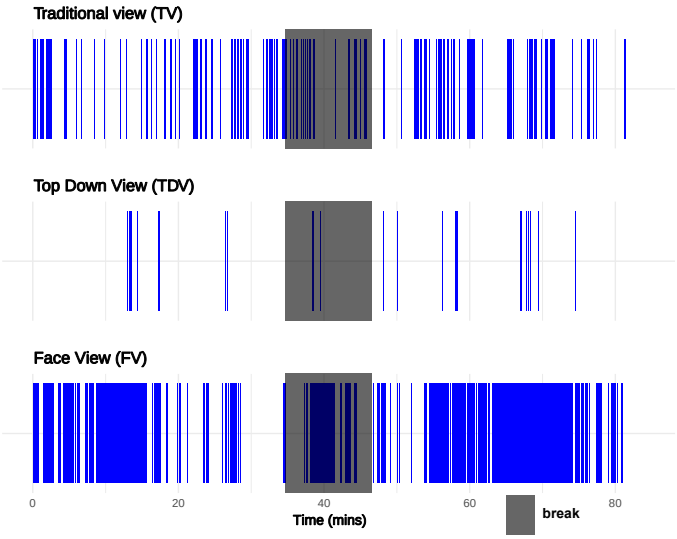

Figure 8: Time series of high quality facial feature detection from all three views

4.3. RESEARCH QUESTION 3

How effectively does OpenPose, a posture estimation toolkit, extract body features from all students simultaneously during small group work in the classroom when using 360-degree cameras versus traditional cameras?

Table 4 highlights a notable disparity in the detection of high-completeness frames detected using OpenPose among the three view angles: TV, TDV, and FV. The FV condition again demonstrates a significantly higher efficacy, yielding 16762 high-completeness frames, which constitutes 14.25% of the total frames detected. Conversely, both TDV and TV showed markedly lower success rates, with only 474 high-completeness frames detected for TDV, amounting to a mere 0.40% of the total frames and 219 high-completeness frames detected for TV, representing 0.18% of the total frames.

The time series plots in Figure 10 showcase the detection of high-completeness body keypoints by OpenPose over the duration of the class session for the three views. The visualizations reveal that although sparse, high-completeness body keypoints were detected for FV at a relatively consistent rate over time when compared to the other views.

Figure 11 presents time series plots illustrating the detection of high-completeness frames for each student across two views: FV and TDV, both captured using 360-degree footage. The color coding differentiates frames with high completeness (blue) from those with low completeness (red), while classroom break is indicated in gray. FV captures more high-completeness frames than TDV across all students. However, the performance of TDV varies significantly between students. For instance, S1 and S2 maintain relatively consistent high-completeness frames across both views, whereas S3 and S4 show a noticeable decline in TDV. In particular, TDV S3 and TDV S4 exhibit frequent instances of low completeness and tracking breaks, suggesting difficulties in detecting body keypoints. This discrepancy is likely due to the inverted orientation of S3 and S4 in the 360-degree footage, which caused keypoint merging and inversion issues during OpenPose's detection. Further elaboration on these challenges is provided in the discussion section.

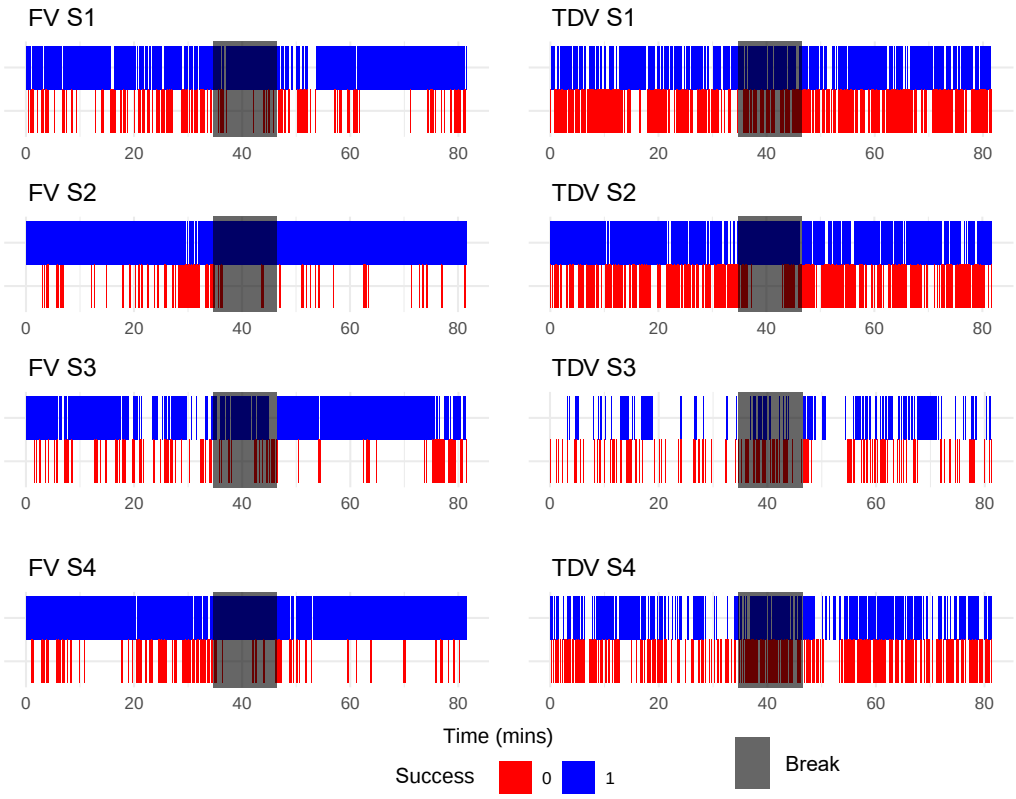

Figure 9: Time series of successful detection of facial features of each student in FV vs TDV

5. Discussion

This exploratory study examined the application of 360-degree camera technology in Multimodal Learning Analytics (MMLA) to enhance the capture and analysis of facial features and body pose estimation in collaborative learning settings. Recognizing the limitations inherent in traditional video capture methods, particularly tripod-mounted video cameras, this research sought to address the challenges posed by these conventional techniques in accurately capturing facial features and body keypoints, which are foundational for building and analyzing collaborative behaviors.

We analyzed three views—Face View (FV), Top-Down View (TDV), and Traditional View (TV)—and assessed their efficiency in detecting facial features and body pose estimations using the OpenFace and OpenPose toolkits. Notably, FV consistently outperformed the other views in terms of efficiency, demonstrating better mean confidence and completeness metrics (see Tables 2 and 3). The alignment of students' faces and bodies toward the 360-degree camera, centrally placed at the table where group interactions occurred, made the FV particularly effective for facial feature and body pose detection. The relatively high percentage of high-quality frames for facial feature detection, coupled with the high-completeness frames that captured all students simultaneously, proved essential for gauging both individual and group-level behaviors (see Table 4). Additionally, FV outperformed the other views in consistently capturing high-quality and high-confidence frames over time (see Figures 8 and 10), which is crucial for developing real-time feedback systems that rely on data quality. These insights are valuable for educa-

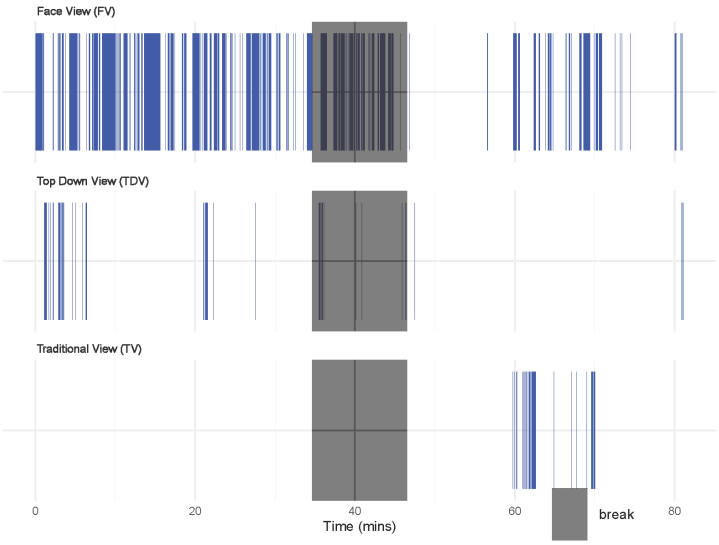

Figure 10: Time series of high-completeness body keypoints detection from all three views

tional researchers seeking to obtain higher-quality datasets to detect meaningful collaborative interaction behaviors within small groups.

However, FV is not without limitations. It encountered issues with the merging of body points across different students near the borders of each student's frame (See Figure 12), as reported in the literature (Cao et al., 2021; Mishra et al., 2024). This issue could potentially be mitigated by utilizing a panoramic view of the 360-degree footage, which provides a wider and more comprehensive view of the classroom. However, even with a panoramic view, there are challenges to address. Specifically, the system needs to ensure that keypoints from students in the background or from other groups, as well as from teachers, are not mistakenly merged with those of the students being tracked (Hur and Bosch, 2022).

The TDV, despite the initial expectation that it would provide a unique bird's-eye view of the group table, capturing inter-student interactions and enabling easy detection of students looking at each other, did not perform to its full potential. TDV struggled to compete with FV in terms of the overall capture of high-quality and high-completeness frames and was less consistent across the duration of the session. This under-performance could be attributed to the significant warping inherent in the bird's-eye perspective. The fact that the OpenFace and OpenPose toolkits had not been trained on fisheye perspective videos likely contributed to the reduced detection accuracy. TDV also encountered issues with the merging of body points across student and background student/instructor (See Figure 12). Additionally, the inverted orientation of students S3 and S4 further complicated the detection process. OpenFace, trained on datasets featuring normally oriented faces, incorrectly applied normally oriented keypoints to the inverted faces (see Figure 13, right). Similarly, OpenPose, trained on datasets with normally oriented bodies (head up and feet down), sometimes flipped shoulder and arm keypoints (Gu et al., 2020; Mishra et al., 2024)(see Figure 13, left). These issues highlight the need for improved algorithms or methods specifically tailored to 360-degree camera perspectives. Despite these challenges, TDV could still be beneficial for analyzing tasks that involve the physical manipulation of materials on the group table, potentially complementing the information obtained from other views.

The TV was the least effective among the three views. The perspective of TV was prone

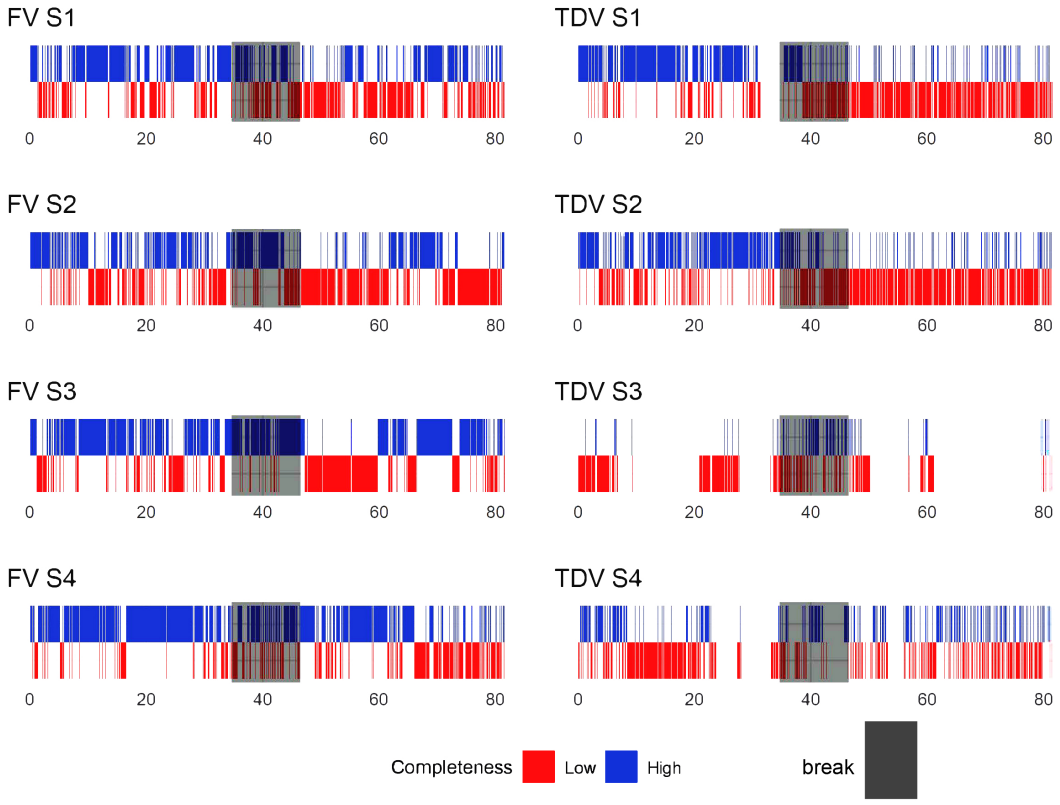

Figure 11: Time series of high-completeness frames for each student in FV vs TDV

to occlusion between students within the group, leading to mis-tracking of facial features and merging of body keypoints (see Figure 12). Occlusion also hindered the tracking of students' bodies on the side opposite the camera's location. As a result, TV's performance was particularly poor for body keypoint detection compared to the other views (see Table 4) as highlighted by researchers (Tsai et al., 2020; Mishra et al., 2024).

A consistent challenge across all perspectives was the management of missing data in time series analysis, a critical factor in ensuring reliable longitudinal analyses. The variability in high-quality, high-completeness frame data necessitates effective strategies for handling missing or incomplete data to ensure the reliability of longitudinal analyses. Approaches such as assigning placeholder values or employing imputation techniques can mitigate the impact of missing data. Recent studies have explored various imputation methods, ranging from statistical techniques like linear interpolation to advanced deep learning-based approaches such as DeepMVI (Bansal et al., 2021) and MultiLayer Perceptron models for filling long continuous gaps (Park et al., 2022). Selecting an appropriate method is context-dependent and should consider data patterns, analysis goals, and computational resources. While our study did not directly address the scalability challenges of applying these techniques across multiple groups in a classroom setting, future work will explore methods to optimize data processing.

As MMLA advances toward the automatic detection of complex latent constructs, such as confusion (Ma et al., 2024), within collaborative learning, it becomes apparent that traditional classroom data collection methodologies must evolve. Integrating high-quality data capture

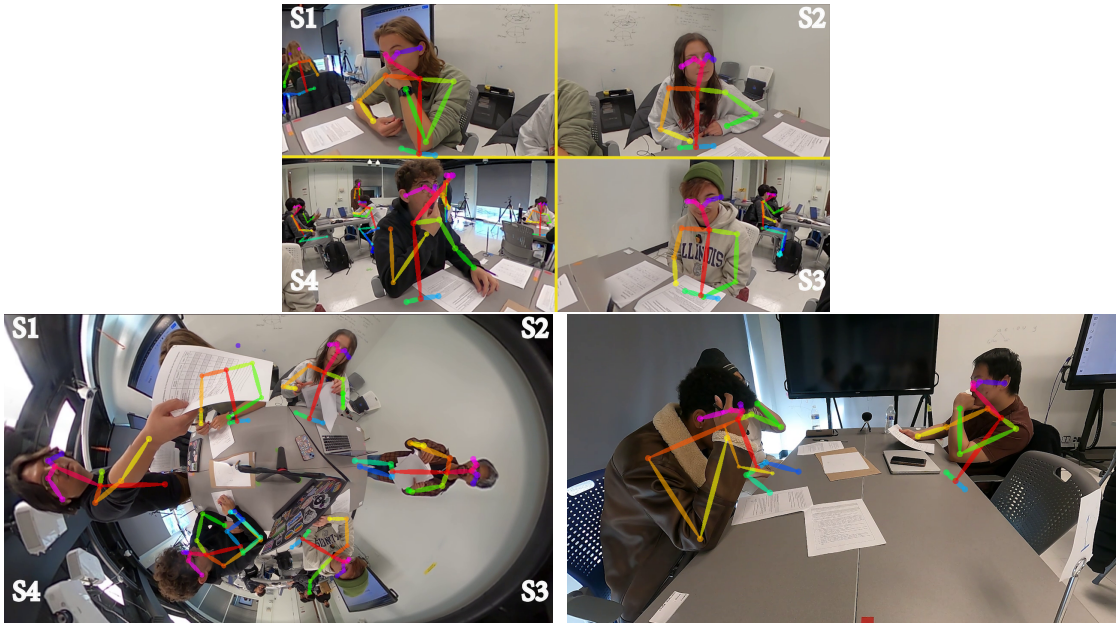

Figure 12: Occlusion and merging issues in all three views when using OpenPose: Merging of S4 with background (FV, top-left), occlusion of S1 and merging of S3 with background (TDV, top-right), and merging students on the left (TV, bottom).

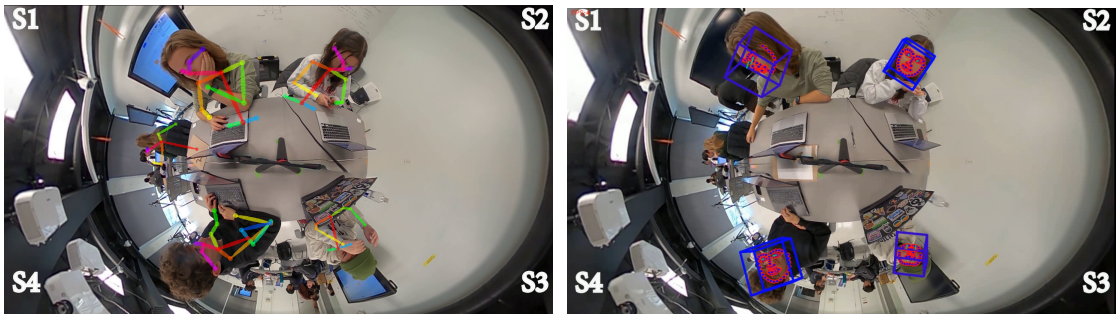

Figure 13: Inversion issues in TDV: OpenPose - S3 (left) and OpenFace - S3 and S4 (right).

tools is paramount for accurately capturing the nuanced behaviors indicative of these constructs. In this light, our study advocates for the use of 360-degree cameras, especially those capable of capturing a 360-degree panoramic view, in face-to-face collaborative activities, particularly those involving groups of three or four students. These devices facilitate comprehensive capture of student-student interactions and enable the extraction of high-quality facial features using toolkits like OpenFace, which provides facial action units (AUs) such as Brow Lowering (AU4), Eyelid Tightening (AU7), and Lip Tightener (AU23)—all essential for affect detection (D'Mello et al., 2009; Padrón-Rivera et al., 2016; Ma et al., 2022). They also enable the capture of high-completeness body keypoints, providing features such as spine similarity, distance between students, and head movement, which are used to identify collaborative behaviors such as turn taking, open to collaboration, closed to collaboration, and synchronized learning (Radu et al., 2020).

However, the adoption of 360-degree cameras is not without challenges. The high cost and

substantial pre-processing requirements pose significant barriers to widespread implementation. Moreover, issues such as keypoint merging and the need for manual tracking highlight technological limitations that need to be addressed. Future developments should focus on refining both hardware and software solutions to overcome these obstacles, potentially making this technology more accessible for large-scale, real-time classroom applications.

6. LIMITATIONS

While this study offers valuable insights into enhancing data capture for collaborative learning analytics, several limitations must be acknowledged. First, the study was conducted with a limited dataset from a specific group and interaction context, which may limit the generalizability of the findings. Variations in classroom settings, group dynamics, and instructional activities could influence the effectiveness of the camera views and detection toolkits. Future research should involve a larger and more diverse sample to validate and extend the applicability of the results.

Second, this study focused primarily on the technical aspects of data capture and analysis using OpenFace and OpenPose toolkits. However, the claims regarding the utility of 360-degree cameras for improving the detection of collaborative behaviors, such as engagement and group interaction patterns, were not empirically validated against a collaborative learning construct. Conducting a follow-up MMLA study specifically targeting a collaborative construct, such as group cohesion or shared understanding, would be necessary to substantiate the proposed benefits of using 360-degree camera technology in educational research.

Third, the limitations associated with the technology itself, including the issues with keypoint merging and the challenges posed by fisheye perspectives, highlight the need for further refinement in both the hardware and software used for such studies.

Finally, the need for manual tracking and the generation of facial view manually needs to be automated to be able to scale this tool for large scale real-time classroom implementation. While 360-degree cameras offer a promising avenue for comprehensive data capture, the high cost and intensive pre-processing requirements remain significant barriers to widespread adoption.

7. FUTURE WORK

Building on the findings of this study, future research will focus on employing 360-degree cameras to explore the affective states and engagement levels of student groups in greater depth. By leveraging facial action units (AUs), head and body posture estimations, and gesture interactions, we aim to analyze these elements in the context of various collaborative constructs, such as group cohesion, shared understanding, and conflict resolution. This will provide a more nuanced understanding of how students interact and collaborate in real-time, contributing to the development of more responsive and adaptive educational tools.

In addressing the challenges of missing data, future studies will explore and compare different imputation methods to determine the most effective approaches for handling incomplete datasets in MMLA contexts. Understanding the minimum data requirements for accurate detection of engagement and affective states, especially in real-time scenarios, will be a key focus.

With the rapid advancement of end-to-end multimodal models, future studies will investigate their performance compared to traditional toolkits like OpenPose and OpenFace in analyzing 360-degree camera footage. Such comparisons will be crucial for determining the efficacy and

potential advantages of adopting newer, integrated approaches within MMLA. These evaluations will focus on aspects such as accuracy, processing efficiency, and ease of use. Incorporating models specifically trained on fisheye and inverted perspectives could address some of the detection issues identified.

Automation of student tracking and isolation of target groups within 360-degree footage is another critical area for future work. Utilizing state-of-the-art tracking tools like the Segment Anything Model (SAM) by Meta AI, which allows for object identification through prompting, could facilitate effective tracking of specific students across videos (Ravi et al., 2024). Tracking students of interest and ignoring background students and teachers is an ongoing issue in the implementation of such tools. Also, SAM's capability to handle partial and temporary occlusions makes it a promising solution for overcoming current limitations.

8. Conclusion

The exploration of 360-degree camera technology in the context of Multimodal Learning Analytics (MMLA) marks a significant advancement in the methodologies used for analyzing collaborative learning processes. This study has shown that 360-degree cameras offer substantial improvements over traditional tripod-mounted cameras, particularly in capturing high-quality facial expressions and high-completeness body postures, which are critical for understanding group dynamics and individual participation within collaborative settings. However, the research also brought to light certain challenges, especially with the Top-Down View (TDV) perspective. The detection of inverted faces and the difficulties in accurately recognizing body keypoints underscore the need for further refinement in both the technology and the algorithms used for facial and pose recognition across different camera views. These findings highlight the necessity of continuing to enhance the tools and approaches used in MMLA to better capture and interpret the complex behaviors that occur during collaborative learning. In addition, the amount of high-quality frames captured suggests the need for researching how to handling missing data while using these tools.

As MMLA continues to evolve, this research underscores the importance of integrating advanced video capture technologies to more accurately assess and interpret the subtle and intricate aspects of collaborative interactions. At the same time, it is essential to acknowledge the practical challenges associated with the adoption of 360-degree cameras, including their cost and the substantial computational resources required for data processing. Addressing these challenges will be crucial for the broader implementation of such technologies in educational research and practice, ultimately contributing to the development of more effective and responsive learning environments.

9. Acknowledgments

We would like to thank our undergraduate research assistant Jacob Frank Sobel for manually verifying how students were tracked from OpenPose data.

This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

DECLARATION OF GENERATIVE AI SOFTWARE TOOLS IN THE WRITING PROCESS

During the preparation of this work, the authors used chatGPT in all the sections in order to improve the grammar and transition between the paragraphs. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

ALWAHABY, H., CUKUROVA, M., PAPAMITSIOU, Z., AND GIANNAKOS, M. 2022. The evidence of impact and ethical considerations of multimodal learning analytics: A systematic literature review. In *The Multimodal Learning Analytics Handbook*, M. Giannakos, D. Spikol, D. D. Mitri, K. Sharma, X. Ochoa, and R. Hammad, Eds. Springer International Publishing, Cham, Switzerland, Chapter 7, 289–325.

AMOS, B., LUDWICZUK, B., AND SATYANARAYANAN, M. 2016. OpenFace: A general-purpose face recognition library with mobile applications. Technical Report CM-CS-16-118, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

ANDRADE, A. 2017. Understanding student learning trajectories using multimodal learning analytics within an embodied-interaction learning environment. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. LAK '17. Association for Computing Machinery, New York, NY, USA, 70–79.

BAKER, R. S. J. D., GOWDA, S. M., WIXON, M., KALKA, J., WAGNER, A. Z., SALVI, A., ALEVEN, V., KUSBIT, G. W., OCUMPAUGH, J., AND ROSSI, L. 2012. Towards sensor-free affect detection in cognitive tutor algebra. In *Proceedings of the 5th International Conference on Educational Data Mining*, K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, and J. Stamper, Eds. International Educational Data Mining Society, Chania, Greece, 126–133.

BALTRUSAITIS, T., ZADEH, A., LIM, Y. C., AND MORENCY, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, IEEE, Xi'an, China, 59–66.

BANSAL, P., DESHPANDE, P., AND SARAWAGI, S. 2021. Missing value imputation on multidimensional time series. *Proc. VLDB Endow. 14,* 11 (July), 2533–2545.

BOTELHO, A. F., BAKER, R. S., AND HEFFERNAN, N. T. 2017. Improving sensor-free affect detection using deep learning. In *Artificial Intelligence in Education*, E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, Eds. Springer International Publishing, Cham, 40–51.

CAO, Z., HIDALGO, G., SIMON, T., WEI, S., AND SHEIKH, Y. 2021. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence 43,* 01, 172–186.

CHEJARA, P., PRIETO, L. P., RODRÍGUEZ-TRIANA, M. J., RUIZ-CALLEJA, A., KASEPALU, R., CHOUNTA, I.-A., AND SCHNEIDER, B. 2023. Exploring indicators for collaboration quality and its dimensions in classroom settings using multimodal learning analytics. In *Responsive and Sustainable Educational Futures: 18th European Conference on Technology Enhanced Learning, EC-TEL 2023, Aveiro, Portugal, September 4–8, 2023, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 60–74.

DAI, Z., MCREYNOLDS, A., AND WHITEHILL, J. 2023. In search of negative moments: Multi-modal analysis of teacher negativity in classroom observation videos. In *Proceedings of the 16th Interna-*

tional Conference on Educational Data Mining, M. Feng, T. Käser, and P. Talukdar, Eds. International Educational Data Mining Society, Bengaluru, India, 278–285.

D'MELLO, S. K., CRAIG, S. D., AND GRAESSER, A. C. 2009. Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology 4,* 3-4, 165–187.

D'MELLO, S. K., OLNEY, A. M., BLANCHARD, N., SAMEI, B., SUN, X., WARD, B., AND KELLY, S. 2015. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI '15. Association for Computing Machinery, New York, NY, USA, 557–566.

EKMAN, P., FRIESEN, W. V., AND HAGER, J. C. 2002. *Facial Action Coding System: The manual: On CD-ROM*. Research Nexus.

EVENS, M., EMPSEN, M., AND HUSTINX, W. 2023. A literature review on 360-degree video as an educational tool: Towards design guidelines. *Journal of Computers in Education 10,* 2, 325–375.

FAHID, F. M., LEE, S., MOTT, B., VANDENBERG, J., ACOSTA, H., BRUSH, T., GLAZEWSKI, K., HMELO-SILVER, C., AND LESTER, J. 2023. Effects of modalities in detecting behavioral engagement in collaborative game-based learning. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. LAK2023. Association for Computing Machinery, New York, NY, USA, 208–218.

GRAESSER, A. C. AND D'MELLO, S. 2012. Moment-to-moment emotions during reading. *The Reading Teacher 66,* 3, 238–242.

GU, Y., ZHANG, H., AND KAMIJO, S. 2020. Multi-person pose estimation using an orientation and occlusion aware deep learning network. *Sensors 20,* 6.

HU, S. AND KUH, G. D. 2002. Being (dis)engaged in educationally purposeful activities: The influences of student and institutional characteristics. *Research in Higher Education 43,* 5, 555–575.

HUR, P. AND BOSCH, N. 2022. Tracking individuals in classroom videos via post-processing openpose data. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. LAK22. Association for Computing Machinery, New York, NY, USA, 465–471.

HUR, P., BOSCH, N., PAQUETTE, L., AND MERCIER, E. 2020. Harbingers of collaboration? The role of early-class behaviors in predicting collaborative problem solving. In *Proceedings of the 13th International Conference on Educational Data Mining*, A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, Eds. International Educational Data Mining Society, Online, 104–114.

LEWIS, A., OCHOA, X., AND QAMRA, R. 2023. Instructor-in-the-loop exploratory analytics to support group work. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference (LAK23)*. ACM, Arlington, TX, USA, 284–292.

LI, Q. AND BAKER, R. 2018. The different relationships between engagement and outcomes across participant subgroups in massive open online courses. *Computers & Education 127*, 41–65.

LIAO, C.-H. AND WU, J.-Y. 2022. Deploying multimodal learning analytics models to explore the impact of digital distraction and peer learning on student performance. *Computers & Education 190*, 104599.

LIU, S., LIU, S., LIU, Z., PENG, X., AND YANG, Z. 2022. Automated detection of emotional and cognitive engagement in mooc discussions to predict learning achievement. *Computers & Education 181*, 104461.

MA, Y., CELEPKOLU, M., AND BOYER, K. E. 2022. Detecting impasse during collaborative problem solving with multimodal learning analytics. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge (LAK22)*. ACM, Online, 45–55.

MA, Y., SONG, Y., CELEPKOLU, M., BOYER, K. E., WIEBE, E., LYNCH, C. F., AND ISRAEL, M. 2024. Automatically detecting confusion and conflict during collaborative learning using linguistic, prosodic, and facial cues. *arXiv preprint arXiv:2401.15201*.

MALLAVARAPU, A., LYONS, L., AND UZZO, S. 2022. Exploring the utility of social-network-derived collaborative opportunity temperature readings for informing design and research of large-group immersive learning environments. *Journal of Learning Analytics 9,* 1, 53–76.

MARTINEZ-MALDONADO, R., ECHEVERRIA, V., FERNANDEZ-NIETO, G., YAN, L., ZHAO, L., ALFREDO, R., LI, X., DIX, S., JAGGARD, H., WOTHERSPOON, R., OSBORNE, A., SHUM, S. B., AND GAŠEVIĆ, D. 2023. Lessons learnt from a multimodal learning analytics deployment in-the-wild. *ACM Transactions on Computer-Human Interaction 31,* 1.

MISHRA, P. K., MIHAILIDIS, A., AND KHAN, S. S. 2024. Skeletal video anomaly detection using deep learning: Survey, challenges, and future directions. *IEEE Transactions on Emerging Topics in Computational Intelligence 8,* 2, 1073–1085.

NOËL, R., MIRANDA, D., CECHINEL, C., RIQUELME, F., PRIMO, T. T., AND MUNOZ, R. 2022. Visualizing collaboration in teamwork: A multimodal learning analytics platform for non-verbal communication. *Applied Sciences 12,* 15.

OVIATT, S., LIN, J., AND SRIRAMULU, A. 2021. I know what you know: What hand movements reveal about domain expertise. *ACM Transactions on Interactive Intelligent Systems 11,* 1 (Mar.).

PADRÓN-RIVERA, G., REBOLLEDO-MENDEZ, G., PARRA, P. P., AND HUERTA-PACHECO, N. S. 2016. Identification of action units related to affective states in a tutoring system for mathematics. *Journal of Educational Technology & Society 19,* 2, 77–86.

PARK, J., MÜLLER, J., ARORA, B., ET AL. 2022. Long-term missing value imputation for time series data using deep neural networks. *Neural Computing and Applications 35*, 9071–9091.

PEKRUN, R. 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review 18*, 315–341.

PRATA, D., D. BAKER, R. S. J., COSTA, E., ROSÉ, C. P., AND CUI, Y. 2009. Detecting and understanding the impact of cognitive and interpersonal conflict in computer supported collaborative learning environments. In *Proceedings of the 2nd International Conference on Educational Data Mining (EDM 2009)*. www.educationaldatamining.org, Cordoba, Spain, 131–140.

RADU, I., TU, E., AND SCHNEIDER, B. 2020. Relationships between body postures and collaborative learning states in an augmented reality study. In *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Springer International Publishing, Cham, 257–262.

RAJARATHINAM, R. J., PALAGUACHI, C., AND KANG, J. 2024. Enhancing multimodal learning analytics: A comparative study of facial features captured using traditional vs 360-degree cameras in collaborative learning. In *Proceedings of the 17th International Conference on Educational Data Mining*, B. PaaÃŸen and C. D. Epp, Eds. International Educational Data Mining Society, Atlanta, Georgia, USA, 551–558.

RAVI, N., GABEUR, V., HU, Y.-T., HU, R., RYALI, C., MA, T., KHEDR, H., RÄDLE, R., ROLLAND, C., GUSTAFSON, L., MINTUN, E., PAN, J., ALWALA, K. V., CARION, N., WU, C.-Y., GIRSHICK, R., DOLLÁR, P., AND FEICHTENHOFER, C. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

RONDA-CARRACAO, M. A., SANTOS, O. C., FERNANDEZ-NIETO, G., AND MARTINEZ-MALDONADO, R. 2021. Towards exploring stress reactions in teamwork using multimodal physi-

ological data. In *Proceedings of the 1st International Workshop on Multimodal Artificial Intelligence in Education (MAIED 2021)*. CEUR Workshop Proceedings, Utrecht, The Netherlands, 48–59.

ROSCHELLE, J. 1992. Learning by collaborating: Convergent conceptual change. *Journal of the Learning Sciences 2*, 235–276.

SCHNEIDER, B., SHARMA, K., CUENDET, S., ZUFFEREY, G., DILLENBOURG, P., AND PEA, R. 2018. Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning 13,* 3 (Sept.), 241–261.

SNELSON, C. AND HSU, Y.-C. 2020. Educational 360-degree videos in virtual reality: A scoping review of the emerging research. *TechTrends 64,* 3, 404–412.

SPIKOL, D., RUFFALDI, E., AND CUKUROVA, M. 2017. Using multimodal learning analytics to identify aspects of collaboration in project-based learning. In *Proceedings of the 12th International Conference on Computer Supported Collaborative Learning (CSCL)*. International Society of the Learning Sciences, Philadelphia, PA, USA, 263–270.

STEWART, A. E. B., KEIRN, Z., AND D'MELLO, S. K. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction 31,* 4, 713–751.

SÜMER, O., GOLDBERG, P., D'MELLO, S., GERJETS, P., TRAUTWEIN, U., AND KASNECI, E. 2023. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing 14,* 2, 1012–1027.

TAYLOR, R. 2016. The multimodal texture of engagement: Prosodic language, gaze and posture in engaged, creative classroom interaction. *Thinking Skills and Creativity 20*, 83–96.

TEDRE, M., TOIVONEN, T., KAHILA, J., VARTIAINEN, H., VALTONEN, T., JORMANAINEN, I., AND PEARS, A. 2021. Teaching machine learning in k–12 classroom: Pedagogical and technological trajectories for artificial intelligence education. *IEEE Access 9*, 110558–110572.

TSAI, Y.-S., HSIEH, Y.-Z., LIN, S.-S., AND CHEN, N.-C. 2020. The real-time depth estimation for an occluded person based on a single image and openpose method. *Mathematics 8,* 8, 1333.

VRZAKOVA, H., AMON, M. J., STEWART, A., DURAN, N. D., AND D'MELLO, S. K. 2020. Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. LAK '20. Association for Computing Machinery, New York, NY, USA, 295–304.

YAN, L., MARTINEZ-MALDONADO, R., CORDOBA, B. G., DEPPELER, J., CORRIGAN, D., NIETO, G. F., AND GASEVIC, D. 2021. Footprints at school: Modelling in-class social dynamics from students' physical positioning traces. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. LAK21. Association for Computing Machinery, New York, NY, USA, 43–54.

ZHOU, Q., BHATTACHARYA, A., SURAWORACHET, W., NAGAHARA, H., AND CUKUROVA, M. 2023. Automated detection of students' gaze interactions in collaborative learning videos: A novel approach. In *Responsive and Sustainable Educational Futures*, O. Viberg, I. Jivet, P. Muñoz-Merino, M. Perifanou, and T. Papathoma, Eds. Springer Nature Switzerland, Cham, 504–517.

ZHOU, Q. AND CUKUROVA, M. 2023. Zoom lens: An mmla framework for evaluating collaborative learning at both individual and group levels. In *Proceedings of the Third International Workshop on Multimodal Immersive Learning Systems (MILeS 2023)*. CEUR Workshop Proceedings, Aveiro, Portugal, 28–35.