Enhancing Multimodal Learning Analytics: A Comparative Study of Facial Features Captured Using Traditional vs 360-Degree Cameras in Collaborative Learning

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

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

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

ABSTRACT

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 sub optimal data quality for facial expressions 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 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 (AU) and facial keypoints. The use of OpenFace revealed that the 360-degree camera captured high-quality facial features in 33.17% of frames, significantly outperforming the traditional method's 8.34%, thereby enhancing reliability in facial feature detection. The findings suggest a pathway for future research to integrate 360-degree camera technology in MMLA. Future research directions involve refining this technology further to improve the detection of affective states in collaborative learning environments, thereby offering a richer understanding of the learning process.

Keywords

facial features, affect detection, computer-supported collaborative learning, multimodal learning analytics

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 and affective 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 [13]. Moreover, while the affective dimension of learning has gained increasing recognition for its impact on educational outcomes, the methodologies for capturing high quality affective features in collaborative settings remain underexplored.

This study is positioned at the intersection of these critical areas of research, aiming to bridge the gap between theoretical advancements in MMLA and their practical application in educational settings. The focus of this exploratory research is to evaluate the efficacy of 360-degree camera technology in enhancing video data quality for face-to-face collaborative learning scenarios. This technology is scrutinized for its potential to offer a more comprehensive capture of the learning environment compared to traditional video methods. Despite the inherent higher costs and potential demands for greater computational power, this study seeks to establish a foundational understanding of how 360-degree cameras could significantly improve the quality of data necessary for effective detection of affective states during small group interactions. By exploring these avenues, the study seeks to contribute to the ongoing discourse on improving the accuracy and reliability of multimodal data analysis in educational research, through novel ways of collecting informative datasets.

2. BACKGROUND

2.1 MMLA in Collaborative Learning

Collaborative learning is a complex sense-making process in which group of students works together to co-construct knowledge via iterative social interactions [**?**]. During group collaboration, students present their ideas, explain to their peers on understanding concepts, solving tasks, and justify ideas in response to questions, challenges, and conflicts. MMLA represents a significant evolution in offering a sophisticated framework for understanding the intricacies of

R. J. Rajarathinam, C. Palaguachi, and J. Kang. Enhancing multimodal learning analytics: A comparative study of facial features capture using traditional vs 360-degree cameras in collaborative learning. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 551–558, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729882

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 [21] to partly [18], or even completely unmediated scenarios [20], encompassing interactions not only between students and teachers [6] but also within the tangible learning environments [22]. 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 [12], cognitive conditions [16], distractions [10], and stress levels [17] 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 [1]. 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 [9]. 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), a cornerstone in affective domain research within MMLA [15]. 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. [7] and operationalized through technologies like the OpenFace library [2]. In the context of MMLA, video data emerges as a critical medium for capturing nonverbal 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 [23]. The sophisticated analysis of video data, including facial expressions and body language, offers insights into the multifaceted affective factors that influence collaborative learning quality.

2.3 Real Classroom Implementation

Implementing MMLA within classroom environments presents a number of challenges that must be navigated carefully. A primary challenge is the inherent complexity and noisiness of data collected in naturalistic settings. These in turn complicates the process of interpreting student interactions. Both practical and logistical hurdles frequently arise when deploying MMLA innovations in real-world educational contexts. Some of these challenges include difficulties in setting up sensing tools for data collection, the increased workload for educators, and ensuring that the sensing tools do not impede on students' learning experiences [13, 4].

One observational study highlights these challenges by employing several machine learning detection algorithms to detect expressions of negative emotions in a classroom [4]. The study particularly utilized OpenFace to detect negative facial expressions from video data. However, the results showed the face detecting tool missed a majority of faces in the dataset. This was attributed to the tool's inadequacy in recognizing faces when not fully visible, leading to missed detection and inaccuracies in interpreting expressions [4]. Such outcomes highlight the necessity for the collection of high-quality, comprehensive multimodal data to improve the reliability and validity of analytic tools for MMLA. In this study, by testing innovative data collection and analysis methods, we aim to provide researchers with practical insights on optimizing the data collection, leading to future improvements to reliability in using MMLA tools.

2.4 360-degree Camera in Education

The advent of 360-degree camera technology introduces a novel dimension to the capture of educational content, leveraging omni-directional or multi-camera systems to record footage from every angle simultaneously. This technology stitches videos together to create a comprehensive spherical view, allowing users to explore the environment in any direction they choose. Such immersive experiences can be accessed through various devices, ranging from computers and smartphones to Head Mounted Displays (HMDs), offering both non-immersive and immersive viewing experiences respectively [19]. Recent reviews have highlighted the application of 360-degree cameras as educational tools, particularly when used in conjunction with HMDs for an immersive learning experience [8].

Despite the growing interest in leveraging 360-degree video for educational purposes, its application within the domain of MMLA remains unexplored. Specifically, there appears to be a gap in the use of 360-degree cameras for capturing high-quality video data in MMLA research, particularly in the context of face-to-face collaborative learning among small groups. Addressing this gap, our study introduces the use of 360-degree cameras as a potential solution to enhance the quality of video data collection in live, collaborative settings involving groups of three to four students. This approach aims to overcome the limitations of traditional video capture methods, providing a more holistic view of the learning environment and facilitating a comprehensive analysis of group interactions.

2.5 Research Questions

To assess the effectiveness of this technological intervention, the research is guided by two critical questions:

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

(RQ 2) How effectively does OpenFace, a facial recognition toolkit, extract facial features from all students simultaneously within small group interactions?

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 Microsoft HoloLens2 via Universal Windows Platform (UWP) within Unity. A new version of HoloOrbits was developed using WebGL via Unity to make the simulation more accessible to students. The goal of the simulation was 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.

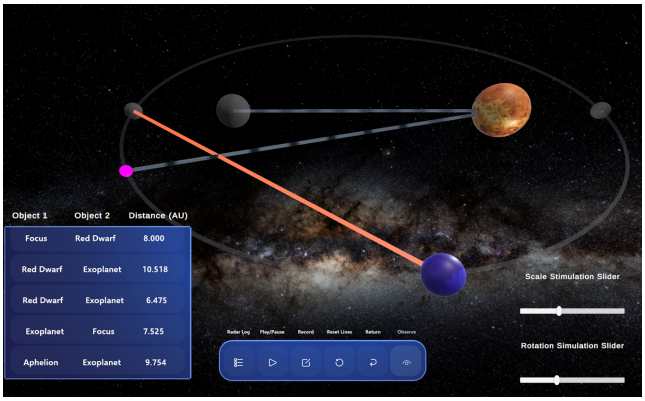

Figure 1: HoloOrbits: A web-based immersive science learning simulation

3.2 Participants

The dataset collected consisted of 2 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 GoPro 360-degree Max camera (Yellow Group) and another group with three students whose video was captured using a tripod and 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, with the red group benefiting from a more controlled viewing angle, as the camera was pointed towards a corner of the classroom.

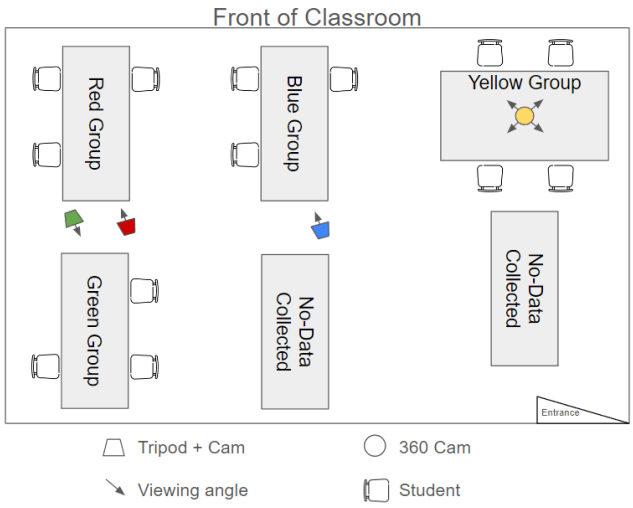

Figure 2: Schematic of the classroom for data collection

3.3 Post-processing

The 360-degree cameras 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 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.

OpenFace v2.2.0 facial behavior analysis toolkit[3] 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

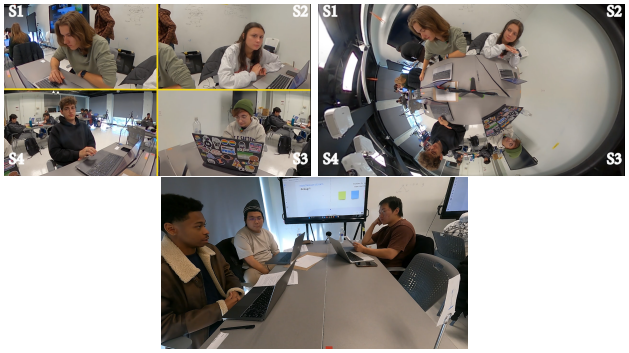

Figure 3: Three video capture views. FV (top-left) and TDV (top-right) are from a 360-degree camera and TV is from a tripod and traditional camera (bottom)

to 0.75 has a success variable (binary) equal to 1. These frames contain facial action units or AUs which are instrumental in detecting affect.

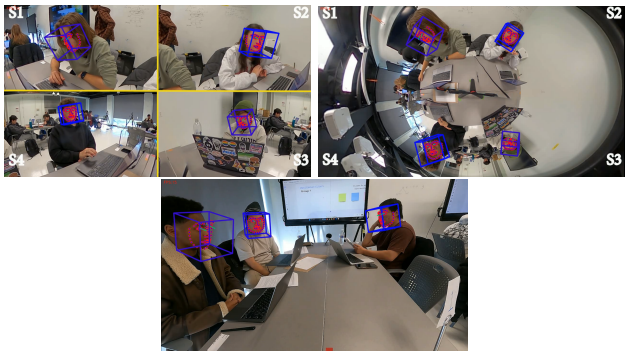

Figure 4: High quality OpenFace facial feature capture. FV (top-left) and TDV (top-right) are from a 360-degree camera and TV is from a tripod and traditional camera (bottom)

3.4 Analysis

In this study, we analyzed a single session to remove external variables like session duration, intervention types, and instructor-student interaction. 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. 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 analysis. Additionally, a Traditional View (TV) was obtained from a conventional video capture setup using a tripod and GoPro Hero10.

Group-wide High quality facial feature detection criteria were 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 facial features in this paper. A confidence threshold of 0.75 was set as a prerequisite for detecting facial action units using OpenFace. This criteria is important in educational research as the lower the detection, the less examples or events can be extrapolated and interpreted. This major increase in high quality facial frames found in FV compared to TV, explains the need for better camera angle capturing for this particular extraction of facial feature data. This stringent selection ensures the reliability of future affect detection in collaborative learning environments, highlighting the significance of AUs captured for all group members.

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, 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 Top-Down, Face Views, and Traditional view, alongside statistical analysis for high-quality video feature capture for the three views, enables a rigorous evaluation of OpenFace—a facial recognition toolkit—in accurately extracting facial features from all students within small group interactions.

4. RESULTS

(RQ 1) Table 1 shows the descriptive statistics of video data from TV, FV, and TDV angles, revealing significant variations in data capture efficiency and processing endpoints. The TV and FV angles demonstrated superior capture efficiency, with unique frame counts significantly exceeding that of the TDV angle, which indicated potential detection limitations or increased frame redundancy.

Table 1: Descriptive statistics summary of video frames by view

View Angles	Total Frames	Unique Frames	Unique Frame Percentage (%)
TV (Traditional)	120,504	115,509	95.85
FV (Face)	117,648	116,535	99.05
TDV (Top-Down)	117,648	106,266	90.31

The histograms in Figure 5 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.

The Kruskal-Wallis test's significant result ($\chi^2 = 82273.378$, $df = 2$, $p = 0$) indicates that there are differences in the median confidence values across the three camera views (FV, TDV, and TV). The Dunn's post-hoc analysis with Bonferroni correction as shown in Table 2 confirms that each pair of groups significantly differs from each other in terms of 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 FV-TDV and FV-TV comparisons sug-

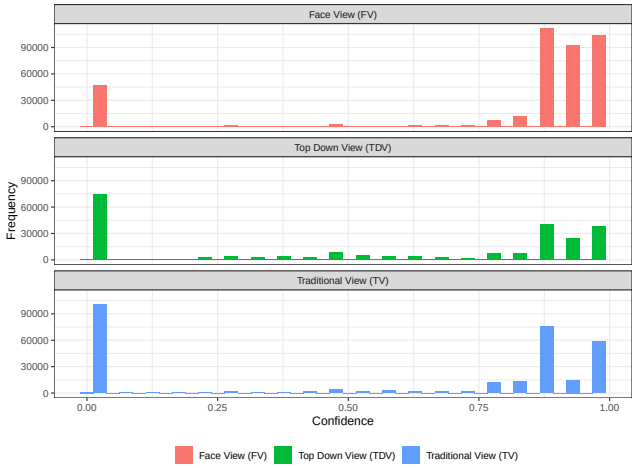

Figure 5: Histograms displaying the distribution of confidence levels with frame frequency across three views

gest higher rank sums (and median confidence) for FV compared to TDV and TV, respectively. The significance across all comparisons suggests that the camera views have statistically significantly different impacts on the confidence values, with FV potentially having the highest median confidence values, followed by TV and then TDV, based on the directionality indicated by the Z-values.

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

Pairwise Comparison	Z-value	p-value	Adjusted p-value
FV vs. TDV	242.35	$< 0.0001^*$	$< 0.0001^*$
FV vs. TV	238.34	$< 0.0001^*$	$< 0.0001^*$
TDV vs. TV	-18.99	$< 0.0001^*$	$< 0.0001^*$

* indicates statistical significance.

Table 3 shows a notable disparity in the detection of high-quality frames 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 analyzed. Conversely, the TDV angle shows a markedly lower success rate, with only 533 high-quality frames recorded, 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 8.34% of the total frames.

Table 3: Number of frames with high quality facial feature detection and their percentage of total frames

View angles	High quality frames	Percentage of total frames (%)
TV (Traditional)	6,215	8.34
TDV (Top-Down)	533	0.45
FV (Face)	39,026	33.17

(RQ 2) The time series plots in Figure 6 showcase the capture of high quality facial features across time 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.

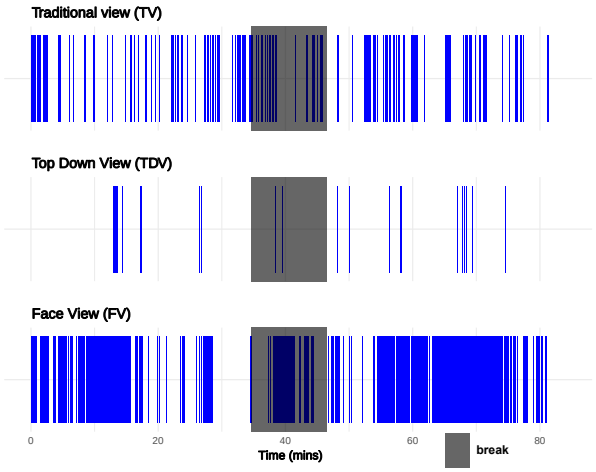

Figure 6: Time series of high quality facial feature detection from all three views

Figure 7 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, FV S1 and TDV S1 showcase differences in the amount of facial features of the frames captured for student 1 (S1). 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.

5. DISCUSSION

This study examined the application of 360° camera technology in MMLA to enhance the capture and analysis of facial features 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 that these conventional techniques pose in accurately capturing facial features, a key element in analyzing collaborative behaviors.

The variation in unique frame detection across camera views indicates their varying efficiency in facial feature capture for data analysis. Top-Down view's (TDV) lower frame count suggests difficulties in facial features captured, whereas Traditional view (TV) had an average performance. Face view (FV) showed the highest efficiency in frames detected for facial features by OpenFace. Focusing on high quality facial feature recognition, the FV outperformed the other views. This high percentage underscores the effectiveness of facial detection technology when applied to face view, likely due to the full visibility of facial features, which facilitates accurate detection and recognition. Additionally, the times series plot of FV shown in 6 showcases the consistently high quality facial feature recognition across time compared to the other views. This may be crucial, especially when building real-time feedback systems which rely on the quality of

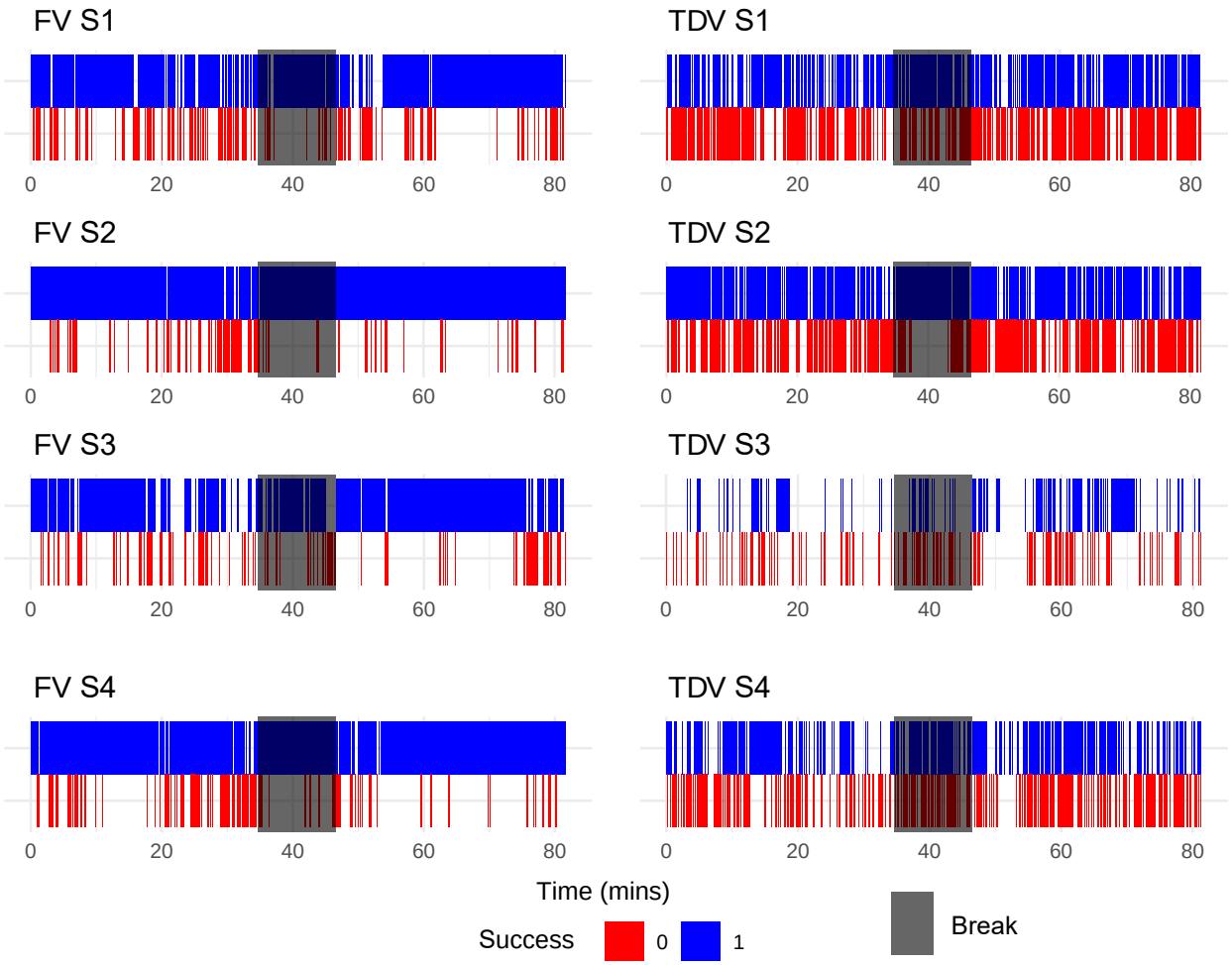

Figure 7: Time series of successful detection of facial features for FV vs TDV

the data collected. These insights could help educational researchers obtain higher quality datasets to detect meaningful collaborative interaction behaviors within small groups.

However, despite being captured from 360-degree camera, TDV struggled in the detection of facial features. Due to the way OpenFace has been trained to recognize human faces, it was particularly bad at detecting faces that were upside down, which is a feature of the TDV setup (see S3 and S4 in TDV in Figure 3). This highlights the need for improved algorithms or methods tailored to 360° camera perspectives. Nevertheless, we remain optimistic that TDV will enhance the ability to capture peer-to-peer posture-based interactions more effectively than the FV.

As MMLA advances towards the automatic detection of complex latent constructs, such as confusion [12], within the realm of collaborative learning, the necessity for evolving traditional classroom data collection methodologies becomes apparent. The integration of high-quality data capture tools is paramount to accurately capturing the nuanced behaviors indicative of these constructs. In this light, this study advocates for the use of 360-degree cameras, such as those with the capability of capturing a 360 panoramic view, for face-to-face collaborative activities, especially involving groups of 3 or 4. These devices not only facilitate comprehensive capture of student-student interactions but also provide the means to extract high-quality facial features using toolkits like OpenFace that provides facial action units (AUs) like Brow Lowering (AU4), Eyelid Tightening (AU7), and Lip Tightener (AU23) which are essential for affect detection [5, 14, 11]. However, it is imperative to acknowledge the inherent challenges associated with the adoption of 360-degree cameras. Despite their potential to offer detailed insights through high-quality facial features, the cost implications and the substantial pre-processing requirements for feature extraction cannot be overlooked. These cameras represent a significant investment and necessitate considerable computational resources for data processing.

Future work will employ 360 cameras to understand the interaction between affective states of student groups using AUs and their collaborative interactions for various collaboration constructs. In addition, the exploration of pose estimators like OpenPose across various camera views to enhance the detection of pose estimation will further support

the robust analysis of collaborative learning environments. Through these efforts, we aim to bridge the gap between MMLA technology and its practical application, contributing to the advancement of educational technologies.

6. CONCLUSION

The exploration of 360-degree camera technology for MMLA represents a pivotal advancement in data collection methods employed for the analysis of collaborative learning processes. The study demonstrated that 360-degree cameras provide a substantial improvement over traditional tripod-mounted cameras in capturing high-quality facial expressions, which are essential for understanding group dynamics. However, the study also highlighted challenges, particularly with the TDV perspective, where the detection of inverted faces proved problematic, indicating a need for a refined approach for better facial and pose recognition across various camera views. As the field of MMLA continues to evolve, this research emphasizes the importance of integrating advanced video capture technologies to accurately assess and interpret the nuanced aspects of collaborative learning, while also acknowledging the practical challenges, such as cost and computational demands associated with 360-degree cameras.

7. ACKNOWLEDGMENTS

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.

8. REFERENCES

[1] H. Alwahaby, M. Cukurova, Z. Papamitsiou, and M. Giannakos. The evidence of impact and ethical considerations of multimodal learning analytics: A systematic literature review. In M. Giannakos, D. Spikol, D. D. Mitri, K. Sharma, X. Ochoa, and R. Hammad, editors, *The Multimodal Learning Analytics Handbook*, chapter 7, pages 289–325. Springer International Publishing, 2022.

[2] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. 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, 2016.

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[4] Z. Dai, A. McReynolds, and J. Whitehill. In search of negative moments: Multi-modal analysis of teacher negativity in classroom observation videos. *International Educational Data Mining Society*, 2023.

[5] S. K. D'Mello, S. D. Craig, and A. C. Graesser. Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology*, 4(3-4):165–187, 2009.

[6] S. K. D'Mello, A. M. Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, and S. Kelly. 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, page 557–566, New York, NY, USA, 2015. Association for Computing Machinery.

[7] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System: the manual: on CD-ROM*. Research Nexus, 2002.

[8] M. Evens, M. Empsen, and W. Hustinx. A literature review on 360-degree video as an educational tool: towards design guidelines. *Journal of Computers in Education*, 10(2):325–375, 2023.

[9] A. C. Graesser and S. D'Mello. Moment-to-moment emotions during reading. *The Reading Teacher*, 66(3):238–242, 2012.

[10] C.-H. Liao and J.-Y. Wu. Deploying multimodal learning analytics models to explore the impact of digital distraction and peer learning on student performance. *Computers & Education*, 190:104599, 2022.

[11] Y. Ma, M. Celepkolu, and K. E. Boyer. Detecting impasse during collaborative problem solving with multimodal learning analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 45–55, 2022.

[12] Y. Ma, Y. Song, M. Celepkolu, K. E. Boyer, E. Wiebe, C. F. Lynch, and M. Israel. Automatically detecting confusion and conflict during collaborative learning using linguistic, prosodic, and facial cues, 2024.

[13] R. Martinez-Maldonado, V. Echeverria, G. Fernandez-Nieto, L. Yan, L. Zhao, R. Alfredo, X. Li, S. Dix, H. Jaggard, R. Wotherspoon, A. Osborne, S. B. Shum, and D. Gašević. Lessons learnt from a multimodal learning analytics deployment in-the-wild. *ACM Trans. Comput.-Hum. Interact.*, 31(1), nov 2023.

[14] G. Padrón-Rivera, G. Rebolledo-Mendez, P. P. Parra, and N. S. Huerta-Pacheco. Identification of action units related to affective states in a tutoring system for mathematics. *Journal of Educational Technology & Society*, 19(2):77–86, 2016.

[15] R. Pekrun. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, 18:315–341, 2006.

[16] D. N. Prata, R. S. d Baker, E. d. B. Costa, C. P. Rosé, Y. Cui, and A. M. De Carvalho. Detecting and understanding the impact of cognitive and interpersonal conflict in computer supported collaborative learning environments. *International Working Group on Educational Data Mining*, 2009.

[17] M. A. Ronda-Carracao, O. C. Santos, G. Fernandez-Nieto, and R. Martinez-Maldonado. Towards exploring stress reactions in teamwork using multimodal physiological data. In *CEUR Workshop Proceedings*, volume 2902, pages 49–60. CEUR-WS, 2021.

[18] B. Schneider, K. Sharma, S. Cuendet, G. Zufferey, P. Dillenbourg, and R. Pea. Leveraging mobile eye-trackers to capture joint visual attention in

co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning*, 13(3):241–261, Sept. 2018.

[19] C. Snelson and Y.-C. Hsu. Educational 360-degree videos in virtual reality: A scoping review of the emerging research. *TechTrends*, 64(3):404–412, 2020.

[20] O. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 14(2):1012–1027, 2023.

[21] H. Vrzakova, M. J. Amon, A. Stewart, N. D. Duran, and S. K. D'Mello. 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, page 295–304, New York, NY, USA, 2020. Association for Computing Machinery.

[22] L. Yan, R. Martinez-Maldonado, B. G. Cordoba, J. Deppeler, D. Corrigan, G. F. Nieto, and D. Gasevic. Footprints at school: Modelling in-class social dynamics from students' physical positioning traces. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, page 43–54, New York, NY, USA, 2021. Association for Computing Machinery.

[23] Q. Zhou and M. Cukurova. Zoom lens: An mmla framework for evaluating collaborative learning at both individual and group levels. In *CEUR Workshop Proceedings*, volume 3499, pages 28–35. CEUR Workshop Proceedings, 2023.